Artificial Intelligence for Bias Detection in Higher Education Online Content

Rawan Mohammed A Bin Shiha

School of Computer Science

University of Leeds

Submitted in accordance with the requirements for the degree of
Doctor of Philosophy

January 2026

Declaration

I confirm that all work presented in this thesis is entirely my own. The publications included were all authored by me as first author. My supervisors are listed as co-authors solely in their roles as academic advisors, providing guidance and oversight, but the research design, analysis and writing were carried out by me.

During the course of this research, I produced five publications, detailed below:

Chapter 4:

1. Bin Shiha, R., Atwell, E. and Abbas, N., 2023, November. Detecting Bias in university news articles: a comparative study using BERT, GPT-3.5 and Google bard annotations. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence* (pp. 487-492). Cham: Springer Nature Switzerland.

2. Bin Shiha, R., Atwell, E. and Abbas, N., 2024, November. University News: A New Data Source for NLP Bias Research. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence* (pp. 307-312). Cham: Springer Nature Switzerland.

Chapter 5:

3. Shiha, R.B., Atwell, E. and Abbas, N., 2022. Decolonising the reading lists of Arabic, Islamic and Middle Eastern Studies. *International Journal on Islamic Applications in Computer Science and Technology*, *11*(2).

4. Shiha, R.B., Atwell, E. and Abbas, N., 2025, November. A Transformer-Based Framework for Thematic Analysis of University Reading Lists. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence* (pp. 405-410). Cham: Springer Nature Switzerland.

Chapter 6:

5. Shiha, R.B., Atwell, E. and Abbas, N., 2025, November. Bias Detection in Online Higher Education Texts: Comparing Fine-Tuned and Zero-Shot LLM Approaches. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence* (pp. 285-290). Cham: Springer Nature Switzerland.

Acknowledgements

I would like to express my deep gratitude to my supervisors, Eric Atwell and Noorhan Abbas, whose insight, guidance and consistent encouragement have been central to the development of this research. I am sincerely appreciative of the time, attention and thoughtful direction they provided throughout this work.

I am grateful to my parents, Norah Bennsar and Mohammed Bin Shiha, for their continuous support and to my siblings for their encouragement. I also appreciate Aunt Hissa Bin Shiha, who has always been there for me.

My heartfelt appreciation extends to my friends, Shahad Battar Alghamdi and Angelo Caldeira, whose constant reassurance and support made the journey lighter. I am also grateful to my dog, Slinky, whose loyal companionship and loving presence offered a calming influence during extended periods of work.

I am profoundly grateful to my country, the Kingdom of Saudi Arabia, for granting me the scholarship that enabled me to pursue my studies in the United Kingdom; my appreciation cannot be fully expressed in words.

Finally, I acknowledge the use of AI tools including Generative Pre-trained Transformer (GPT) and Claude.ai's Sonnet 4.5 which supported aspects of proofreading and clarity during the preparation of this thesis.

Abstract

This thesis develops and evaluates Artificial Intelligence (AI) and Natural Language Processing (NLP) approaches for detecting bias in higher education online resources, addressing the lack of systematic computational methods in this area. Bias in higher education content shapes knowledge production, representation and equity, making its detection both academically and socially significant.

To address this gap, three sets of novel datasets were created: 1. a corpus of university news articles annotated for subjectivity, sentiment and gender representation; 2. three university reading list datasets with demographic annotations enabling comparative analysis across Western and Middle Eastern contexts; and 3. two domain-specific collections of learning materials, one from humanities-oriented open resources and the other from Science, Technology, Engineering and Mathematics (STEM) lecture transcripts. Together, these datasets provide the first systematic resources for investigating representational, stereotypical and linguistic bias across diverse higher education domains.

Using these resources, a range of Pre-trained Language Models (PLMs) and Large Language Models (LLMs) were evaluated for bias detection. PLM revealed significant gendered and representational disparities in university discourse, while fine-tuned LLM achieved improved performance on humanities data but showed limited transferability to STEM materials. A hybrid framework integrating fine-tuned LLMs with Retrieval-Augmented Generation (RAG) enhanced detection transparency and prediction balance. These methods were operationalised in a prototype web application for bias detection in higher education learning content.

The contributions of this research are fourfold: 1. the creation of multiple novel annotated datasets spanning university news, reading lists and academic learning resources; 2. the introduction of LLM-based strategies for bias annotation, fine-tuning and cross-domain evaluation 3. methodological innovations including a structured framework for categorising bias, a hybrid human–AI annotation approach and a replicable NLP pipeline for demographic and thematic analysis; and 4. the design of a hybrid bias detection system and accompanying web application.

This work advances computational approaches to bias detection, provides reproducible resources for future research and offers practical tools to support greater fairness and equity in higher education online environments.

Abbreviations

AI Artificial Intelligence

NLP Natural Language Processing

STEM Science, Technology, Engineering and Mathematics

PLMs Pre-trained Language Models

LLMs Large Language Models

RAG Retrieval-Augmented Generation

BAME Black, Asian and Minority Ethnic

ML Machine Learning

NLI Natural Language Inference

NER Named Entity Recognition

LDA Latent Dirichlet Allocation

CNN Convolutional Neural Networks

SVM Support Vector Machines

DT Decision Trees

RF Random Forests

ANN Artificial Neural Networks

UoL-NA University of Leeds News Articles

AIMESRL Arabic, Islamic, and Middle Eastern Studies Reading List

QSIERL Qur'an Sciences and Islamic Education Reading List

DSRL Disability Studies Reading List

OER Open Educational Resources

OER-LMH Open Educational Resources Learning Materials in Humanities

UoL-DSPDS-MLT University of Leeds Data Science and Programming for Data Science Modules Lecture Transcripts

MTurk Amazon Mechanical Turk

API Application Programming Interfaces

MultiNLI Multi-Genre Natural Language Inference

SD Standard Deviation

FRE Flesch Reading Ease

UMAP Uniform Manifold Approximation and Projection

HDBSCAN Hierarchical Density-Based Spatial Clustering of Applications with Noise

c-TF-IDF Class-based Term Frequency Inverse Document Frequency

ARI Adjusted Rand Index

TLS Transport Layer Security

AWS Amazon Web Services

GPT Generative Pre-trained Transformer

UI User Interface

HTTPS Hypertext Transfer Protocol Secure

PII Personally Identifiable Information

WEAT Word Embedding Association Test

MAB Mean Absolute Bias

MDB Maximum Difference Bias

SLMs Smaller Language Models

CGANs Conditional Generative Adversarial Networks

UMLS United Medical Language System

GenAI Generative Artificial Intelligence

MAIC Massive AI-empowered Course

MOOCs Massive Open Online Courses

LC-RAG Log-Contextualised RAG

NWSA National Woman Suffrage Association

LMS Learning Management Systems

Contents

List of Figures

List of Tables

Chapter 1

1 Introduction

1.1 Research Gap, Objectives, Aim and Contributions

The fields of NLP and AI have increasingly focused on the critical challenge of bias detection in textual data. Work in this area includes Iyyer et al. (2014), who developed methods for detecting political bias in text, whilst Hube and Fetahu (2018) made a contribution through their detection of biased statements in Wikipedia. Cruz et al. (2020) built systems capable of detecting and classifying biased news articles. Doughman and Khreich (2022) released labelled datasets and lexicons for multiple gender-bias subtypes to facilitate automated detection of gender bias in English text.

With the rapid development of LLMs capable of generating coherent and contextually rich text, researchers have begun to use these models both as tools and as objects of study to investigate bias in text. Soundararajan and Delany (2024) use LLMs to generate gendered sentences and assess bias via downstream stereotype classification and lexical analyses of adjective usage. Dong et al. (2023) probe both explicit and implicit gender bias through conditional text generation from LLMs. Masoudian et al. (2025) analyse LLM-generated stories to observe how gender contributions shift under conditioning.

Building on these efforts to detect and analyse bias in text, scholars have sought to categorise the many ways such bias can manifest. Bias in textual data manifests in multiple forms, each presenting unique detection challenges. Gender bias, which has received considerable attention in NLP research, can appear through gendered language, stereotypical role assignments or unequal representation of different genders (Sun et al., 2019; Stanczak and Augenstein, 2021). Sun et al. (2019) identify four forms of representation bias in NLP models, noting that these systems often propagate and may even amplify gender biases found in training corpora. Ethnic bias emerges through the marginalisation of particular ethnic groups, the perpetuation of stereotypes, or the dominance of certain racial perspectives (Davidson et al., 2019; Sap et al., 2019). Beyond gender and ethnic biases, textual corpora investigated in NLP bias research have been shown to contain political bias (Iyyer et al., 2014), religious bias (Abid et al., 2021), disability bias (Hutchinson et al., 2020) and socioeconomic bias (Field, Anjalie et al., 2021) and more.

Despite major advances in bias detection across news, social media and other textual domains, gender and ethnic bias in higher education materials remains critically underexplored. Existing research on curricula and reading lists in higher education highlights a persistent lack of representation, with university materials still dominated by Eurocentric and male authors (Phull et al., 2019; Schucan Bird and Pitman, 2020). This underrepresentation contributes to feelings of exclusion among students from minority backgrounds (AbouElMagd, 2016). While campaigns such as "Why is My Curriculum White?"

(National Union of Students 2019), #LiberateMyDegree (Bhambra et al., 2018) and "Decolonising the Curriculum" (Hall et al., 2021) have called attention to these issues, there is still limited empirical work that systematically analyses bias in the actual text of higher education resources, especially as these materials increasingly move online. This gap leaves educators and institutions without the evidence base needed to develop more inclusive, representative digital learning environments.

As higher education increasingly moves online, these issues extend beyond traditional printed textbooks and classroom syllabi. Many universities have also begun to routinely upload recorded lectures, lecture slides and transcripts to their online platforms to support flexible learning, a practice accelerated by the shift to remote teaching during the COVID-19 pandemic (Voelkel et al., 2023). Moreover, student reliance on university online resources surged during and following the pandemic (Howard et al., 2023).

A smaller but growing body of work has begun to investigate bias in digital educational content more broadly. For example, Gezici and Saygin (2022) analyse YouTube educational videos and find a significant bias toward male narrators in STEM vs non-STEM video search results. Santos et al. (2022) develop a method to detect gender-based colour bias on educational technology web pages. A recent study by Khokhlova et al. (2023) uses an experimental design with online lectures to show the effect when participants evaluate male vs female virtual instructors. However, despite these valuable contributions, none of these studies focus on bias within the text of online higher education materials, leaving a critical gap that this research seeks to address.

Table 1.1 presents selected examples of biased text identified in this research within higher educational content. These examples were extracted from online higher education resources in Humanities disciplines such as history, law and social sciences, including lecture transcripts, notes, essays, teaching guides and related materials. For further information on bias categorisation and its analytical framework, see Chapter 3, Section 3.2, and for details regarding the data, see Section 3.3.

Example	Bias Type	Bias Category	Context	Why is this bias?
"Students should be able to respond effectively to the following questions: What was happening to the right to vote in the first half of the 19th century? Is it likely that the newly enfranchised voters would have differences in concerns from those who had already been voting? ... Help the students think about the likely differences in the concerns of each that might affect voting, if s/he had the franchise: A New Jersey widow whose husband left her a small fortune and a successful shipbuilding business..."	Gender	Representation Stereotype	This was extracted from a teaching guide that asks students to assess early-19th-century voting rights changes expanding male suffrage after 1800 and the brief New Jersey women's franchise (1776-1807).	This exemplifies stereotype and representation bias: the guide portrays the woman solely as "a New Jersey widow whose husband left her a small fortune and a shipbuilding business" defining her identity through her spouse. This representation reduces her to a derivative figure, obscuring independent political or social agency.
"The National Woman Suffrage Association (NWSA) soon rallied around a new strategy: the New Departure. This new approach interpreted the Constitution as already guaranteeing women the right to vote... In 1875, the Supreme Court addressed this constitutional argument: acknowledging women's citizenship but arguing that suffrage was not a right guaranteed to all citizens... Following this defeat, many suffragists like Stanton increasingly replaced the ideal of universal suffrage with arguments about the virtue that White women would bring to the polls. These new arguments often hinged on racism and declared the necessity of White women voters to keep Black men in check."	Gender	Stereotype Representation Linguistic	This passage is from an educational learning platform covering the history of the women's suffrage movement in the United States. While providing historical context about the NWSA and its legal strategies.	This segment contains bias in its description of how the suffrage movement shifted its arguments. The text frames White women as inherently virtuous voters, reinforcing gender stereotypes about female moral superiority. Additionally, the phrase "White women voters to keep Black men in check" perpetuates racist stereotypes by portraying Black men as a threat requiring control.

"…During the following class session give the principal speaker for each side an allotted amount of time to make his or her speech. Do the same for the second speakers (usually less time than the first). Then throw the debate open so that team members from each side can question or make comments to the other side..."	Gender	Linguistic	This passage is from an educational activity guide that instructs teachers on how to conduct a classroom debate.	This text demonstrates gender bias through its use of "his or her" which reflects a binary understanding of gender that excludes non-binary individuals. More inclusive language, such as 'their' or 'the speaker' would better acknowledge the full spectrum of gender identities.
"Give your evaluation of the strongest claims Malcolm X made regarding why integration will not work and justify your choice. Make sure you include a consideration of why some blacks would have viewed his approach as legitimate in the early 1960s. Do you think Malcolm X gave sufficient evidence to support the claim that black Americans deserve a separate nation? Why does Malcolm X disagree with both the goal and the method of Martin Luther King, Jr.'s nonviolent protest movement? How did Martin Luther King, Jr. expect that his nonviolent protest methods would help to produce a "beloved community"? Connect his means with his ends."	Ethnic	Representation Linguistic	This is an essay prompt from a higher education political science course examining the Civil Rights Movement.	The phrasing promotes a competitive comparison that implies one leader's philosophy is superior while also demonstrating ethnic representation bias by subjecting Malcolm X's ideas to greater scrutiny than Martin Luther King Jr.'s, positioning integrationist approaches as more legitimate. This asymmetrical framing oversimplifies their complex perspectives and reinforces a hierarchy among Black political thought rather than treating both as equally valid responses to systemic racism.
"We hold as undeniable truths that the governments of the various States, and of the confederacy itself, were established exclusively by the white race, for themselves and their posterity; that the African race had no agency in their	Ethnic	Stereotype Representation Linguistic	This passage is from Texas's Declaration of Causes for secession (February 2, 1861), one of several declarations issued by Southern states to justify leaving the Union. In higher	This text promotes a racial ideology legitimising slavery through claims of white superiority and divine sanction. It describes African people as inherently inferior and asserts that slavery benefits both enslaved and

establishment; that they were rightfully held and regarded as an inferior and dependent race... That in this free government all white men are and of right ought to be entitled to equal civil and political rights; that the servitude of the African race, as existing in these States..."			education, these declarations are studied as primary historical sources in courses on American history, Civil War studies, and critical race theory to examine the ideological foundations of the Confederacy.	enslavers by God's will, manipulating religious narratives to justify racial hierarchies while portraying opposition as catastrophic.
"The Federal Government... has for years almost entirely failed to protect the lives and property of the people of Texas against the Indian savages on our border, and more recently against the murderous forays of banditti from the neighboring territory of Mexico... When we advert to the course of individual non-slave-holding States... our grievances assume far greater magnitude."	Ethnic	Linguistic Stereotype	This passage is from Texas's Declaration of Causes for secession (February 2, 1861). The document uses dehumanising language such as "Indian savages" to describe Native Americans and "banditti" for Mexicans, employing ethnic stereotyping to portray non-white populations as threatening and uncivilised while justifying Texas's grievances against the Federal Government. In higher education, this document is analysed as a primary source in American history and ethnic studies.	This text employs racially charged language that perpetuates stereotypes and dehumanises marginalised groups. The term "Indian savages" stereotypes and dehumanises Native Americans, while "murderous forays of banditti" criminalises Mexicans, fostering prejudice.

Table 1.1: Illustrative examples of bias in online higher education Humanities content

The examples presented in Table 1.1 require careful contextualisation. It is important to clarify that the identification of bias within historical texts does not equate to a dismissal of their historical significance or educational value. Rather, historical documents are particularly valuable sites for examining how language reflects and reinforces dominant ideologies, power relations and social hierarchies. The presence of racially charged terms such as "Indian savages" and "banditti" illustrates how linguistic choices were used to construct narratives of fear, superiority and justification for exclusion or violence. In the context of this research, the purpose of highlighting such language is not to

anachronistically judge the past, but to demonstrate how AI and NLP methods can systematically detect patterns of dehumanisation, stereotyping and rhetorical framing that continue to influence the interpretation of educational content. Recognising bias within historically significant material therefore enhances, rather than diminishes, its pedagogical relevance by enabling more critical and reflective engagement with the past.

Further examples of bias in higher educational resources examined in this research include gender representation in university news articles. An analysis of gender distribution at the University of Leeds, UK, revealed notable disparities in the use of male and female terms. Of the 5,768 articles analysed, 34% predominantly featured male-related terms, 12% featured female-related terms and the remaining 52% were classified as neutral, containing no predominant gender terms (see Chapter 4, Section 4.4.4).

An additional example of bias was also identified in university module reading lists, specifically in relation to author representation. An analysis of 212 reading list items from modules within the School of Arabic, Islamic and Middle Eastern Studies at the University of Leeds, UK, showed that male authors accounted for 85% of the dataset, with 80% of these male authors being Western (see Chapter 5, Section 5.3).

Addressing bias in text within online higher education materials is crucial for promoting equitable and inclusive learning environments. Evidence from social psychology shows that representation of diverse voices in educational content fosters belonging and engagement, particularly for students from historically marginalised groups (Steele and Aronson, 1995; Walton and Cohen, 2011). Many Black, Asian and Minority Ethnic (BAME) students report feeling underrepresented and finding their perspectives inadequately reflected in mainstream academic narratives (AbouElMagd, 2016; Schucan Bird and Pitman, 2020). Without intervention, biased curricula risk reinforcing structural inequalities, reproducing stereotypes and discouraging participation from underrepresented students (Delgado and Stefancic, 2023).

As universities increasingly adopt digital approaches, the reach and influence of online resources expand significantly, amplifying the impact of any bias they contain (Munoz-Najar et al., 2021). Addressing these biases is not only an educational priority but also a computational one. Most benchmarks for NLP bias detection focus on news or social media corpora rather than educational text, leaving a gap in domain-specific evaluation.

Therefore, this research addresses that gap by analysing a novel combination of online higher education materials that have not previously been examined together for gender and ethnic bias detection. The study draws on three distinct data sources:

- British university news articles.
- Universities modules reading lists from both British and Saudi universities.
- Domain-specific higher education learning materials (Humanities disciplines texts and STEM-focused lecture transcripts).

To the best of our knowledge, there is no dataset on bias in university news articles, no resource that explores reading list author diversity through cross-cultural comparison, and no annotated dataset for gender and ethnic bias in higher education learning resources. This study presents the first datasets addressing all three of these gaps

The aim of this research is to design and evaluate NLP and AI approaches for detecting gender and ethnic bias in higher education online resources. the study pursues the following objectives:

- Develop a structured framework for categorising bias in online higher education resources.
- Build and annotate diverse datasets from online higher education content to enable bias detection.
- Applies and compares a range of NLP and AI approaches for bias in higher education content.
- Create a prototype web-based tool to translate the research findings into a practical system for bias detection in online educational materials.

This research therefore contributes both to the education sector and to NLP bias detection research. For education, it offers the first systematic, data-driven examination of bias in online higher education resources. For NLP, it expands the scope of bias detection into a critical but underexplored domain, enabling models to operate more fairly in educational contexts.

Detecting gender and ethnic bias in the selected materials is a challenging task requiring linguistic nuance and advanced computational modelling. Bias cues are often subtle, context-dependent and vary across domains (Sap et al., 2019; Sheng et al., 2019; Blodgett et al., 2020). University news articles may convey bias through framing, lecture transcripts through conversational discourse and reading lists through patterns of inclusion or omission. These differences mean a single uniform approach is insufficient, requiring domain-sensitive NLP methods that can adapt across heterogeneous sources.

1.2 Research Questions

The main research question

How can AI and NLP driven frameworks be designed, evaluated and deployed to detect bias across diverse higher education online resources?

Supporting research questions

1. What key types of gender and ethnic bias (e.g., stereotype) are present in higher education online resources, how can these be defined and operationalised for detection?

2. To what extent can LLMs based bias annotation and investigation pipelines effectively capture different manifestations of bias across varied higher education online resources?

3. How effective are zero-shot PLMs, LLMs and fine-tuned LLMs as NLP and AI tools for generating content and detecting bias in higher education online sources?

4. How can the RAG method be integrated into a bias detection framework for higher education online resources and what advantages does it offer in terms of accuracy, robustness and transparency?

5. How can a bias detection framework be implemented in a web application tool to make bias identification methods more accessible and usable within academic environments?

1.3 Contributions

This thesis makes several contributions to the study of bias detection in higher education online resources.

1.3.1 Bias Classification in Online Higher Education Resources

Introduce a structured framework for defining and categorising bias across diverse online higher education resources.

1.3.2 Novelty in Data

1.3.2.1 University News Articles Dataset

A corpus of university news articles categorised each news article based on the presence of gender identity terms. Articles were categorised as male (M), female (F) or neutral (N) depending on whether they predominantly featured masculine or feminine possessive nouns or neither. Moreover, subsets of this data are annotated for subjectivity bias and sentiment. This dataset provides the first systematic resource for investigating gendered discourse in university reporting.

1.3.2.2 University Reading List Datasets

Three curated datasets containing demographic annotations for author gender and ethnicity. Two datasets enable comparative studies of representational bias and curricular diversity across Western and Middle Eastern contexts, while a third captures patterns of diversity within interdisciplinary settings.

1.3.2.3 Higher Educational Learning Materials Datasets

Two domain-specific datasets of higher education learning resources. The first is drawn from an online learning platform covering Humanities disciplines, while the second consists of university lecture transcripts in STEM disciplines. These datasets extend bias detection research into formal academic materials across contrasting domains.

1.3.3 LLM Bias Annotation, Fine-tuning and Investigation

1.3.3.1 LLM-Driven Bias Annotation Strategies

Introduce LLM-based bias annotation strategies for different online higher educational resources.

1.3.3.2 Domain-Specific Bias Detection LLM

Fine-tune a domain-specific bias detection LLM tailored to Humanities disciplines.

1.3.3.3 Cross-Domain Bias Transfer Evaluation

Conduct a cross-domain transfer experiment by applying an LLM trained on humanities data to STEM data, evaluating its capacity to detect bias across distinct disciplinary contexts.

1.3.3.4 Hybrid LLM-RAG Framework for Bias Detection

Design a hybrid framework integrating fine-tuned LLM model with RAG for bias detection.

1.3.4 Methodological Contributions

1.3.4.1 Hybrid Human-LLM Annotation Method

Develop a hybrid annotation methodology combining human expertise with LLM-assisted annotation.

1.3.4.2 NLP Pipeline for Reading-List Diversity

Design a replicable NLP pipeline to analyse demographic and thematic diversity in university reading lists.

1.3.4.3 Gender Bias Detection in University News

Initiate the first NLP-based research study to detect gender bias in university news articles.

1.3.4.4 Web Prototype for Bias Detection

Create a web application prototype for bias detection in Humanities disciplines within higher education.

1.4 Thesis Outline

Before outlining the structure of this thesis, it is important to consider the methodological approach underpinning this research. This research employs a deliberately progressive approach to methodological complexity, beginning with foundational techniques and advancing towards sophisticated AI systems. For example, the study initially utilises zero-shot classification methods before introducing fine-tuning approaches, allowing for a comparative evaluation of baseline and enhanced model performance. This strategy ensures robust validation at each stage, enabling the research to assess how incremental improvements in methodology contribute to more accurate and nuanced bias detection.

This progressive methodology was particularly necessary given the pioneering nature of this research. As one of the first studies to investigate bias specifically within higher educational texts, there were few established frameworks or prior methodologies to guide the approach. The limited body of existing literature in this domain presented considerable challenges, from sourcing appropriate data to selecting suitable analytical methods. To our knowledge, no existing dataset specifically addresses bias detection within higher educational content, nor has any prior textual dataset been developed using higher educational materials for bias detection purposes. This absence required the careful curation and preparation of original data sources. The lack of precedent also meant that method selection demanded extensive experimentation and iterative refinement. These challenges further justify the progressive research design, which allowed for flexibility and adaptation as new insights emerged.

A further contribution to the novelty of this research is its use of three distinct data sources, providing a more comprehensive approach to investigating bias in higher educational content. The research employs a deliberately designed multi-modal data collection strategy that systematically examines bias across different levels of educational discourse. Unlike previous studies that examine bias within a single dataset in a domine or a single data source of at a time (Shah, D. et al., 2019; Blodgett et al., 2020), this work adopts a triangulated approach that captures bias manifestations across institutional

communications, curricular materials and pedagogical content. This methodological innovation addresses a critical gap identified by Larson et al. (2017) and Davidson et al. (2019), who argued that bias detection systems trained on single domains often fail to generalise across educational contexts due to domain-specific linguistic patterns and cultural framings.

However, the use of three distinct datasets, whilst strengthening the comprehensiveness of the research, also added considerable complexity to the methodological process. Each dataset presented unique characteristics, structures and linguistic features, requiring tailored approaches for data preparation, analysis and bias detection. Identifying methods capable of effectively handling this diversity proved particularly challenging, as techniques suitable for one data type did not always transfer successfully to others. This necessitated extensive testing, comparison and refinement of multiple approaches, which inevitably slowed the research progress. Furthermore, the lack of existing benchmarks or comparative studies meant that evaluating the effectiveness of chosen methods required additional validation steps. Despite these obstacles, the decision to utilise multiple data sources was essential to ensure the findings reflect the multifaceted nature of bias within higher educational contexts.

Ultimately, the scarcity of prior work in this area, combined with the comprehensive multi-modal approach adopted, highlights both the difficulty and the significance of this contribution to the field.

The thesis is organised into four sections, beginning with the theoretical background and moving through dataset creation, model-based investigations and practical web tool.

1.4.1 Section 1: Foundations

- **Chapter 2:** This chapter reviews the theoretical and methodological foundations of the research. It introduces key concepts of bias in digital and academic contexts; surveys prior work in NLP for bias detection and discusses the challenges specific to higher education online resources.

1.4.2 Section 2: Datasets and Annotation

- **Chapter 3:** This chapter defines the main categories of bias relevant to higher education resources and presents the construction of novel datasets across multiple domains. It introduces annotation frameworks to create gold-standard corpora for subsequent analysis.

1.4.3 Section 3: Model-Based Investigations

- **Chapter 4:** This chapter provides the analysis of bias in university news. It evaluates PLM and LLMs for bias subjectivity, identifies gender-based patterns of representation and sentiment detection.

- **Chapter 5:** This chapter investigates demographic and topical diversity in university reading lists across Western and Middle Eastern contexts. Transformer-based NLP methods are applied to detect representational bias, epistemic diversity and institutional differences in curricular design.
- **Chapter 6:** This chapter examines fine-tuned models for bias detection in academic learning resources. It evaluates their strengths and limitations in identifying explicit and implicit bias, compares their performance with baseline methods, and explores cross-domain transferability between humanities and STEM content.
- **Chapter 7:** This chapter introduces a hybrid approach that integrates RAG with fine-tuned LLM model. It analyses how combining these methods influences prediction consistency, interpretability and system behaviour in the context of bias detection.

1.4.4 Section 4: Application and Conclusion

- **Chapter 8:** This chapter translates the research findings into practice through the design and implementation of a web application using one data from the three studies in this thesis. The tool operationalises the bias detection framework, making it usable and accessible within academic environments.
- **Chapter 9:** This chapter summarises the thesis findings, highlights key contributions and reflects on limitations. It concludes with directions for future research.

Chapter 2

2 Background

2.1 Bias in NLP

The concept of bias in NLP systems has been defined in numerous ways across the literature, reflecting different disciplinary perspectives and analytical priorities. Understanding these varied definitions is crucial to identify and mitigate bias. This section examines the major definitions and sources of bias proposed in the literature, the types of bias that have been identified and provides a critical analysis of how these conceptualisations shape bias research.

2.1.1 Conceptualising Bias in NLP: The Social Foundations of Bias

The foundation of bias in NLP lies in recognising that language itself is a social construct, shaped by power dynamics, historical inequalities and societal norms (Guo et al., 2024). When NLP systems inherit and amplify societal biases, they do so because they have learned from training data that reflects these inequalities. As an example, research has demonstrated that LLMs, trained on vast datasets collected from the internet, inevitably inherit and amplify the societal biases embedded in their training data (Ranjan et al., 2024). This process is not accidental; rather, it represents the mechanisation of existing social hierarchies through computational means.

Garrido-Muñoz et al. (2021) provide a foundational sociological definition, stating that bias is "a prejudice in favour or against a person, group or thing that is considered to be unfair". This definition emphasises the social dimensions of bias, rooting it in questions of fairness and justice rather than purely technical concerns. The definition captures bias as a form of prejudice that operates through favouritism or discrimination, affecting individuals or groups in ways deemed unjust. In contrast, the Machine Learning (ML) and statistics literature often conceptualises bias in technical terms as systematic error in measurement or prediction, or as deviation from ground truth in model predictions (Mikołajczyk-Bareła and Grochowski, 2025). This definition focuses on accuracy and model performance rather than social harm, treating bias as a correctable technical problem rather than a reflection of social inequality.

Blodgett et al. (2020) conducted one of the most comprehensive surveys of bias in NLP, analysing 146 papers and identifying a critical problem in how bias is conceptualised. They found that "their motivations are often vague, inconsistent and lacking in normative reasoning, despite the fact that analysing 'bias' is an inherently normative process". This critique reveals that researchers frequently invoke bias without clearly articulating what

system behaviours are harmful to whom, why they are harmful or the normative reasoning underlying these assessments. Blodgett et al. argue that addressing bias requires researchers and practitioners to "articulate their conceptualisations of 'bias' i.e., what kinds of system behaviours are harmful, in what ways, to whom and why, as well as the normative reasoning underlying these statements". This definition reframes bias not as a measurable quantity but as an inherently value-laden concept requiring explicit normative justification.

Gallegos et al. (2024) define social bias as "disparate treatment or outcomes between social groups that arise from historical and structural power asymmetries". This definition explicitly acknowledges that bias is not simply individual prejudice or statistical error but rather emerges from historical and structural dimensions of inequality. The emphasis on "power asymmetries" positions bias as fundamentally about relationships of domination and subordination between groups, moving beyond individual attitudes to encompass systemic patterns of inequality. This definition connects bias in NLP systems and dataset used to train these systems to broader social structures of oppression and privilege.

As an example of the link between the social aspect of bias and NLP applications, Bolukbasi et al. (2016) define bias in word embeddings as undesirable associations captured in vector representations, such as when science-related terms are more strongly associated with male attributes relative to female attributes or when certain ethnic names are more associated with pleasant or unpleasant words. This embedding-based definition treats bias as geometric relationships in high-dimensional spaces, measurable through vector distances and projections. Contextualised embedding approaches extend this notion to sentence-level representations, defining bias as differential associations between social group terms and neutral concepts when embedded in similar linguistic contexts (Kurita et al., 2019).

Another example of the connection between social aspects of bias and NLP systems, from a fairness perspective, has been described as a violation of fairness criteria. Gallegos et al. (2024) articulate multiple fairness desiderata, the violation of which constitutes bias. Fairness through unawareness is violated when "a social group is explicitly used" in model processing, such that outputs depend on protected attributes. Invariance is violated when outputs for analogous inputs differing only in social group membership are not "identical under some invariance metric" meaning the model produces systematically different outputs for different groups given equivalent inputs. Equal social group associations are violated when "a neutral word is not equally likely regardless of social group" such as when occupation terms are predicted with different probabilities depending on gender context. Equal neutral associations are violated when "protected attribute words corresponding to different social groups are not equally likely in a neutral context" meaning the model treats different demographic groups differently in contexts where group membership should be irrelevant.

Understanding the social aspects of bias is crucial in any research involving data collection and analysis. Bias is shaped by cultural, demographic and contextual factors which can significantly influence how data is produced, interpreted and ultimately used. This

research places particular emphasis on identifying and addressing social dimensions of bias during the earliest stages of the process specifically in selecting and annotating data. By focusing on these steps, the study highlights how social context affects the creation and labelling of datasets, which in turn impacts the accuracy and fairness of analytical tools.

2.1.2 Taxonomies of Bias: Types, Sources and Dimensions

The literature reveals multiple types of bias organised according to different taxonomic frameworks. Gallegos et al. (2024) provide the most detailed taxonomy, distinguishing between allocational harms and representational harms as two fundamental categories. Allocational harms are defined as "disparate distribution of resources or opportunities between social groups". Within allocational harms, two distinct types are identified. Direct discrimination is defined as "disparate treatment due explicitly to membership of a social group" such as when LLM-aided résumé screening systems perpetuate existing inequities in hiring by explicitly favouring certain demographic groups. Indirect discrimination is characterised as "disparate treatment despite facially neutral consideration towards social groups, due to proxies or other implicit factors" as occurs when LLM-aided healthcare tools use variables that serve as proxies for demographic factors, thereby exacerbating existing inequities in patient care even without explicitly considering protected attributes.

Representational harms, the second major category in Gallegos et al. (2024) taxonomy, encompass the "perpetuation of denigrating and subordinating attitudes towards a social group". This broad category includes six distinct subtypes. Derogatory language refers to "pejorative slurs, insults or other words or phrases that target and denigrate a social group" such as when language models generate or fail to flag slurs that express contempt towards particular groups. Disparate system performance involves "degraded understanding, diversity or richness in language processing or generation between social groups or linguistic variations" exemplified when African American English expressions are misclassified as invalid English more frequently than Standard American English equivalents. Exclusionary norms represent "reinforced normativity of the dominant social group and implicit exclusion or devaluation of other groups" such as when language like "both genders" implicitly excludes non-binary gender identities. Misrepresentation is defined as "an incomplete or non-representative distribution of the sample population generalised to a social group" occurring when systems respond to statements like "I'm an autistic dad" with expressions of sympathy, thereby conveying negative and incomplete representations of autism. Stereotyping encompasses "negative, generally immutable abstractions about a labelled social group" exemplified when models associate "Muslim" with "terrorist" thereby perpetuating harmful violent stereotypes. Finally, toxicity refers to "offensive language that attacks, threatens or incites hate or violence against a social group" such as expressions of hate directed at specific ethnic or racial groups.

An alternative taxonomic approach focuses on the sources of bias in NLP systems. Hovy and Prabhumoye (2021) identify five distinct sources from which bias emerges in the development pipeline. Data bias arises from training corpora that contain biased content or

are drawn from non-representative samples of the population. Annotation bias is introduced through labelling processes when human annotators apply their own biases in creating training data. Model bias emerges from architectural choices and training procedures, such as optimisation functions that prioritise accuracy over fairness. Research design bias stems from decisions about problem formulation, which applications to develop and how to frame research questions. Evaluation bias results from the selection of benchmarks and metrics that may themselves embody particular values or fail to capture relevant harms. This source-based taxonomy emphasises that bias is not monolithic but enters systems through multiple pathways across the development lifecycle.

Mehrabi et al. (2021) and Suresh and Guttag (2021) extend the analysis of bias sources by examining how bias manifests at different stages. In training data, bias may emerge from historical bias embedded in the world that is reflected in datasets, representation bias when certain groups are underrepresented or overrepresented, measurement bias when proxies are used that imperfectly capture concepts of interest and aggregation bias when data from distinct groups is combined inappropriately. In the model itself, bias may be amplified beyond what exists in training data through learning algorithms and evaluation procedures that optimise for particular objectives whilst ignoring fairness considerations. In evaluation, bias manifests through benchmark datasets that are unrepresentative of deployment populations and through metric choices that obscure disparate performance. In deployment, bias emerges when models are used in contexts different from their training environment or when model outputs are presented in ways that affect human perception and decision-making.

Bias has also been categorised according to the social dimensions or protected attributes it targets. Nangia et al. (2020) developed the CrowS-Pairs dataset and identified nine distinct types of social bias: race, gender, sexual orientation, religion, age, nationality, disability, physical appearance and socioeconomic status/occupation. Several of these correspond directly to the "protected characteristics" set out in UK anti-discrimination law under the Equality Act 2010, namely age, disability, gender reassignment, marriage and civil partnership, pregnancy and maternity, race, religion or belief, sex and sexual orientation (Equality and Human Rights Commission, 2021). In this framework, the CrowS-Pairs categories of race and nationality map onto the protected characteristic of race, which expressly includes colour, nationality and ethnic or national origin. The gender category aligns with the protected characteristic of sex and may intersect with gender reassignment were bias concerns trans identities. The categories of sexual orientation, religion and age each have exact equivalents in the protected characteristics of sexual orientation, religion or belief and age, while disability directly matches the protected characteristic of disability. By contrast, physical appearance and socioeconomic status/occupation are not themselves protected characteristics under the Equality Act 2010, although bias on these grounds can intersect with, or compound, discrimination linked to one or more protected characteristics in practice. Smith, E.M. et al. (2022) expand this categorisation in their HolisticBias framework to encompass 13 demographic axes with nearly 600 associated descriptor terms, reflecting growing recognition that bias operates across multiple intersecting dimensions of identity. These dimension-based taxonomies organise bias

according to which social groups are affected, though they often treat dimensions as separate when real-world bias frequently operates at the intersection of multiple identities.

2.1.3 Manifestations of Bias Across Different NLP Tasks

The literature distinguishes types of bias according to the NLP tasks in which they manifest. In text generation, Liang et al. (2021) and Sheng et al. (2019) identify that bias appears both locally through word-context associations such as different next-token likelihoods for "The man was known for" versus "The woman was known for" and globally through properties of entire generated spans such as overall sentiment or regard towards different groups. Sheng et al. (2019) introduced the concept of "regard" to measure perceptions towards different demographic groups in generated text. In machine translation, Měchura (2022) documents how exclusionary norm bias manifests when translators default to masculine forms in cases of ambiguity, such as translating English "I am happy" to masculine French "je suis heureux" rather than feminine "je suis heureuse".

Vanmassenhove et al. (2019) and Prates et al. (2020) demonstrate that commercial translation systems yield masculine defaults far more frequently than expected from demographic data. In information retrieval, Rekabsaz and Schedl (2020) document how bias emerges through disparate performance in document ranking, with queries returning more documents related to masculine concepts than feminine ones despite gender neutral search terms. In question answering, Dhamala et al. (2021) and Parrish et al. (2021) show that stereotyping bias appears when models rely on group-based generalisations to answer questions in ambiguous contexts, such as assuming certain occupations or behaviours are associated with particular demographic groups. In Natural Language Inference (NLI), Dev et al. (2021) demonstrate how misrepresentation and stereotyping biases cause models to make invalid inferences, such as predicting that "the accountant ate a bagel" entails "the man ate a bagel" rather than maintaining neutrality. In coreference resolution, Rudinger et al. (2018) and Zhao, J. et al. (2018) document how bias manifests when models link pronouns to stereotypically associated occupations connecting "she" with "nurse" and "he" with "engineer" rather than relying on syntactic or contextual information.

2.2 Ethnic and Gender Biases in Educational Texts

The examination of bias in educational materials has evolved considerably over recent decades. Understanding this progression is essential for contextualising contemporary research and appreciating how methodological advances have expanded our capacity to identify and analyse bias in higher education texts.

2.2.1 Background on Ethnic and Gender Bias in Education

Early investigations into bias in educational materials employed traditional qualitative and quantitative methods that, whilst labour-intensive, provided foundational insights into the nature and extent of bias. Sadker, M.P. and Sadker (1980) pioneering content analysis of 24 preservice teacher-education textbooks revealed that less than 1% of content addressed sexism, highlighting significant underrepresentation of women and prevalence of

male-oriented language. This manual approach established important baseline understandings of how gender bias manifested in educational texts. Zittleman and Sadker (2002) revisited this issue two decades later, evaluating 23 teacher education textbooks published between 1998 and 2001 using a 72-item assessment tool. Their findings indicated that gender issues remained segregated into specific chapters rather than integrated throughout texts, suggesting that progress in addressing bias had been incremental despite growing awareness.

Blumberg (2008) comprehensive international review of gender bias in textbooks demonstrated the global nature of this issue, synthesising research from multiple countries that documented consistent patterns of underrepresentation and stereotypical portrayals of women. Alrabaa (1985) primary analysis of Syrian textbooks revealed how males were depicted as dominant characters whilst women were portrayed in subservient roles, directly affecting students' perceptions of gender roles. In China, a Ford Foundation-funded research initiative systematically investigated gender representation in primary and junior secondary school textbooks, identifying widespread patterns of male dominance in textual and visual depictions as well as the reinforcement of traditional gender roles (Blumberg, 2007). Studies of Chinese textbooks revealed systematic gender bias: Ping and Weiling (2002) found that males comprised nearly two-thirds of those depicted in elementary mathematics books and 74% of those shown in stimulating activities, whilst Jin (2002) documented that 100% of scientists and soldiers were portrayed as male. Ldli and Zhenzhou (2002) analysed elementary language textbooks and found females made up only one-fifth of historical characters portrayed, consistently depicted with stereotypical attributes. These patterns persisted into junior middle school materials, where Lili (2003) and Zhao, P. (2003) found contradictions between students' lived experiences of working mothers and the traditional domestic roles portrayed in textbooks.

The movement towards decolonising curricula in higher education has brought ethnic and racial bias into sharper focus. Phull et al. (2019) and Schucan Bird and Pitman (2020) examined university reading lists in the UK, documenting the dominance of white, Western male authors. Their work emerged from student movements such as 'Why is My Curriculum White?' and #LiberateMyDegree (Bhambra et al., 2018), which challenged Eurocentric perspectives in higher education. These studies employed mixed methods including statistical analysis of author demographics and qualitative interviews with students, revealing that BAME students often felt underrepresented and marginalised by curricula that inadequately reflected diverse perspectives. In African contexts, research has documented similar patterns whilst also highlighting efforts to integrate local epistemologies and challenge colonial knowledge structures. Gyamera and Burke (2018) and Laakso and Hallberg Adu (2024) explored how African universities are working to decolonise their curricula, incorporating indigenous knowledge systems alongside Western academic traditions. The #FeesMustFall and #RhodesMustFall movements in South Africa (le Grange, 2016; Hlatshwayo and Fomunyam, 2019) demonstrated the urgency with which students and academics have sought curricular transformation, though the research also indicates that implementing such changes remains complex and contested.

The application of intersectionality theory (Crenshaw, K.W., 1989) to educational texts has enriched understanding of how multiple forms of bias interact and compound. Collins (2022) and Ong et al. (2011) explored the experiences of women of colour in academic settings, particularly in STEM fields, documenting how these students navigate overlapping forms of discrimination. Sesko and Biernat (2010) analysed psychology textbooks, finding stereotypical portrayals of Black women that reflected intersecting racial and gender biases. Moore (2017) examined assessment practices, identifying differential feedback patterns for female students of colour. Amutah et al. (2021) revealed concerning representations of minority patients in medical textbooks, demonstrating how bias in educational materials can potentially influence future professional practice. These studies highlight the importance of considering how various forms of bias operate simultaneously, though they also indicate the methodological challenges inherent in capturing intersectional dynamics. Read et al. (2005) study of gender bias in essay assessment found that whilst subjective grading variations existed, the relationship between assessor characteristics and bias was more complex than initially anticipated, suggesting that institutional and disciplinary norms play significant roles.

The theoretical framework of stereotype threat (Steele and Aronson, 1995) provides a mechanism for understanding how biased texts might affect student performance and engagement. Research by Schucan Bird and Pitman (2020) documented that students exposed to non-diverse reading lists reported feelings of underrepresentation, whilst studies across various contexts (Mlama, 2005; Alayan et al., 2012; Jha et al., 2019) have explored how bias affects educational experiences in Africa, Asia, and the Middle East. These investigations have established important connections between textual representation and student perceptions, though further research is needed to establish direct causal relationships between specific biases and measurable learning outcomes.

2.2.2 Examining Ethnic and Gender Bias in Education Through NLP

The advent of NLP and ML has transformed the scale and scope of bias detection research. Lucy et al. (2020) pioneered the application of NLP techniques to analyse 15 history textbooks used in Texas schools, employing WordNet and spaCy's Named Entity Recognition (NER) tagger alongside Latent Dirichlet Allocation (LDA) for topic modelling. Their computational approach enabled systematic quantification of representation patterns, revealing significant underrepresentation of Latin American individuals and demonstrating how different demographic groups were discussed in varying contexts. This study marked a significant advancement in bias research, as computational methods could process larger text corpora more consistently than manual review, though the authors acknowledged that automated approaches might not capture all forms of subtle or context-dependent bias.

Orgeira-Crespo et al. (2021) extended these computational approaches by employing Convolutional Neural Networks (CNN) to analyse over 100,000 doctoral theses from Spanish universities, identifying patterns of gender bias related to authors' characteristics and academic disciplines. The scale of their analysis would have been impractical using

traditional methods, demonstrating how ML can uncover systemic patterns across extensive datasets. Kausar et al. (2023) compared multiple ML algorithms such as Support Vector Machines (SVM), Decision Trees (DT), Random Forests (RF) and Artificial Neural Networks (ANN) in detecting gender bias, finding varying accuracy rates that highlighted both the potential and limitations of automated approaches. Their study revealed that whilst ANN achieved 87.2% accuracy, different algorithms identified somewhat different instances of bias, raising important questions about validation and ground truth in automated bias detection.

Albuquerque (2023) developed an integrative methodology that combined manual annotation by 193 higher education students with ML models including SVM, RF and XGBoost. This approach recognised that bias perception varies across individuals with different ethnic backgrounds and that automated detection benefits from human judgement to provide contextual understanding. The study found that titles and academic disciplines influenced bias detection, suggesting that context matters significantly in identifying bias. By training ML models on manually annotated data, the research demonstrated how automated and traditional approaches can complement each other, with human insight informing algorithmic detection whilst automation enables broader application.

2.3 LLMs Applications for Higher Education

LLMs are rapidly becoming central to AI for education, moving beyond generic chatbots towards integrated systems for tutoring, learner modelling, feedback and large-scale online teaching (see Table 2.1 for illustrative examples).

Application area	Key examples and findings
Conversational course assistants	Liu, C. et al. (2025) deployed an LLM-based course assistant used in around 2,000 students across six computer science courses at three institutions. The assistant provided on-demand help with course content and programming tasks; log data showed heavy usage in evenings and nights and especially in introductory courses, suggesting that LLMs can address "temporal support gaps" when human office hours are unavailable.
Teaching and assessment support	Bektik et al. (2024) reported, based on a review of 112 sources, that Generative Artificial Intelligence (GenAI) tools like ChatGPT support educators by automating content creation, generating customised lesson plans, providing instant feedback, and enabling more interactive learning environments, particularly in STEM and healthcare disciplines. This can free staff time for higher-value activities such as mentoring and formative dialogue.
Large-scale online education content	Yu et al. (2024) proposed the concept of Massive AI-empowered Course (MAIC), which uses LLM-driven multi-agent systems to create AI-augmented classrooms that combine scalability with adaptivity. The project aims to develop an open platform for LLM-based online education, bringing together educators and researchers. In a preliminary deployment at Tsinghua University, involving over 500 students and more than 100,000 learning records, MAIC agents provided personalised support within large online courses, enabling richer interaction than traditional Massive Open Online Courses (MOOCs).

Table 2. 1: Representative applications of LLMs in higher education settings and their key findings

LLMs are already being applied across a wide range of higher education activities and the examples shown in Table 2.1 represent only a subset of their potential. As the technology matures and institutional policies evolve, we can expect LLM-driven solutions to become

increasingly embedded in teaching, learning, administration and scholarly practice, shaping the future landscape of higher education.

2.4 LLMs for Bias Detection in Text

Systematic approaches have been developed to evaluate biases embedded within LLMs, recognising this as essential for building fair and reliable tools. These methodologies employ diverse evaluation techniques, often categorised by the level at which they operate, such as examining word embeddings, output probabilities and generated text. Increasingly, LLMs themselves are being utilised as tools for bias identification and analysis. An example of this is the work of Albuquerque et al. (2025), which uses LLMs to examine large corpora of educational materials for latent stereotypes or discriminatory patterns.

2.4.1 Bias Embedded in LLM Outputs and Mechanisms

The use of LLMs for bias detection represents a critical intersection of computational linguistics and social equity concerns. A fundamental challenge in this domain is recognising that LLMs, despite their sophisticated language processing capabilities, are not neutral analytical tools. The sources of bias in LLM outputs are multifaceted, operating at multiple stages of model development and deployment (Lee, J. et al., 2024). The origins of bias in LLMs are deeply rooted in training data. Biases in the outputs of generative AI reflect the inherent biases present in massive textual corpora including news items, Wikipedia, books and internet content (Bozkurt et al., 2024).

Since large-scale generative AI is trained on vast amounts of data, it becomes increasingly challenging to monitor or meticulously filter such biases. Consequently, these models tend to inherit and amplify prevailing opinions and stereotypes present during model training, leading to the production of biased content. Studies have shown that LLMs exhibit fundamental social identity biases, characterised by in-group solidarity (i.e., favouring one's own group) and out-group hostility (i.e., derogating other groups), often matching patterns observed in human social psychology (Hu, T. et al., 2025). For instance, investigations involving 77 different LLMs demonstrated that nearly all base models and some instruction-tuned models displayed clear ingroup-positive and outgroup-negative associations when prompted to complete sentences (Hu, T. et al., 2025).

Lee, J. et al. (2024) emphasise that current measures of bias from traditional ML fail to transfer directly to LLM-generated text, such as tutoring conversations, because text encodings are high-dimensional, multiple correct responses may exist, and tailoring responses may be pedagogically desirable rather than inherently unfair. This distinction is crucial for understanding how bias manifests and should be evaluated in educational contexts.

Ranjan et al. (2024) provide a crucial foundational resource by conducting a comprehensive survey on the pervasive issue of biases embedded within LLMs. This work places the models themselves under scrutiny, aiming to delineate the current landscape of these

biases by providing an extensive review of their types, sources, impacts and mitigation strategies. The core methodological output of this work is the systematic categorisation of biases into several dimensions, synthesising extant research findings regarding their manifestation and effects in real-world applications. Specifically, the authors unify the literature by proposing two intuitive taxonomies for bias evaluation metrics and datasets. The first taxonomy organises metrics based on the different levels at which they operate in a model as presented in Table 2.2. This taxonomy also helps to disambiguate the relationship between metrics and evaluation datasets. The second taxonomy for bias evaluation categorises datasets by their structure, primarily as counterfactual inputs or prompts, and identifies the targeted harms and social groups, as shown in Table 2.3.

Category of Bias	Type of Bias	Description
Application-Specific Bias	Task Bias	Occurs when the model performs differently across tasks or domains.
Cognitive Bias	Confirmation Bias	Favouring information that confirms existing beliefs or assumptions.
Data-Driven Bias	Data Bias	Results from skewed or unrepresentative training data.
Data-Driven Bias	Content Bias	Involves biased narratives or topics in the training data.
Data-Driven Bias	Label Bias	Caused by inconsistent or biased labelling during training.
Model Architecture Bias	Algorithmic Bias	Arises from model design or learning algorithms that affect fairness.
Social Bias	Gender Bias	When the model perpetuates gender-based stereotypes or inequities.
Social Bias	Racial Bias	When outputs reflect racial stereotypes or discriminatory views.
Social Bias	Age Bias	Reflects age-related stereotypes in model behaviour or representation.
Societal Bias	Social Bias	Mirrors broader societal prejudices present in data.
Societal Bias	Cultural Bias	Misrepresentation or underrepresentation of cultural groups.
Systemic Bias	Feedback Loop Bias	Reinforcement of bias through user interaction and model feedback.

Table 2. 2: Categories and types of biases in LLMs (Adapted from Ranjan et al. (2024), p.6).

Source of Bias	Description	Example
Training Data	Biases from unrepresentative or skewed training data.	Data from social media reflecting societal prejudices.
Model Architecture	Design decisions that lead to biased behaviour or outputs.	Neural network configurations amplifying certain features.
Human Annotation	Labelling inconsistencies introducing bias.	Annotators inconsistently labelling gender or race.
User Interactions	Feedback loops reinforcing existing biases.	Users favouring biased responses during model training.
Societal Influences	Social norms and stereotypes shaping data and outputs.	Reinforcement of gender roles in generated text.
Prompt Design	Structural bias in how prompts are phrased.	Certain prompt wording leading to skewed responses.
Evaluation Metrics	Metrics that fail to account for fairness or representation.	Relying on accuracy without measuring fairness across groups.
Cultural Context	Lack of cultural awareness in model training.	Overlooking regional language or cultural variations.
Selection Bias	Unequal representation of groups in data.	Overrepresentation of Western contexts.

Table 2. 3: Sources and examples of bias in LLMs (Adapted from Ranjan et al. (2024), p.7).

Huang, Y. (2024) investigated gender bias in LLM-generated teacher evaluations in higher education, specifically focusing on evaluations produced by GPT-4. This research employed a comprehensive analytical framework that included the Odds Ratio analysis, the Word Embedding Association Test (WEAT), sentiment analysis and contextual analysis. The results identified gender-associated language that reflected common societal stereotypes, noting that words related to approachability and support were used more frequently for female instructors, while terms associated with entertainment were predominantly used for male instructors, aligning with communal versus agentic behavioural stereotypes.

Warr et al. (2024) examination of bias in LLMs revealed racial bias in ChatGPT's evaluation of student writing. By manipulating racial descriptors in prompts, the researchers assessed differences in scores given by two ChatGPT models. Their findings indicated that descriptions of students as Black or White led to significantly higher scores compared to race-neutral or Hispanic descriptors, suggesting that ChatGPT's outputs are influenced by racial information in ways that raise substantial concerns about its application in educational assessment contexts.

Additionally, Weissburg et al. (2024) conducted a large-scale evaluation examining how LLMs generate and select educational content tailored to different demographic groups. Utilising two bias score metrics, Mean Absolute Bias (MAB) and Maximum Difference Bias (MDB), the researchers analysed nine open and closed state-of-the-art LLMs across over 17,000 educational explanations at multiple difficulty levels and topics. Their analysis revealed significant biases across race, ethnicity, sex, gender, disability status, income and national origin dimensions. Notably, the models potentially harm student learning by both perpetuating harmful stereotypes and reversing them. The highest MAB appeared along income levels, whilst MDB was highest relative to both income and disability status. For both metrics, the lowest bias existed for sex/gender and race/ethnicity dimensions.

2.4.2 Research Utilising LLMs as Tools for Bias Investigation

The primary application of LLMs in educational settings as bias research tools involves their utilisation in auditing and detecting existing bias within large corpora of text. For example, Albuquerque et al. (2025) work in assessing inter-group bias within learning texts from 91 higher education courses at The Open University UK using LLMs. Three different LLMs produced remarkably different results: BERT identified bias in 49.5% of sentences, compared to 10.2% for GPT and 11.0% for PaLM. Agreement levels between model-to-model and model-to-human comparisons ranged from 0 to 0.77 Kappa values, underscoring the complexity of inter-group bias where context and culture prove crucial. This variation suggests that no single LLM should be considered a reliable arbiter of bias without contextual verification and human judgement. Consequently, Albuquerque et al. (2025) advocate for detection tools that are both context- and culture-aware when identifying inter-group bias in learning materials, recognising that the subtle nature of such bias cannot be adequately captured through purely algorithmic means. Their research emphasises that assumptions about bias detection must account for the nuanced ways that context and cultural frameworks shape the meaning and acceptability of language in educational settings.

Beyond educational contexts, research has employed LLMs as analytical instruments to investigate biases across multiple dimensions. In hate speech detection tasks, researchers have utilised LLMs including GPT-3.5, GPT-4o, Llama-3.1 and Gemma-2 to annotate datasets and subsequently analyse the biases present in their classifications across gender, race, religion and disability categories (Das et al., 2024). By deploying LLMs as annotation tools, researchers were able to systematically examine how these models introduce biases

when labelling highly vulnerable groups within these demographic categories. Similarly, persona-based LLMs have been employed as detection instruments to analyse biases in hate speech annotation tasks, examining how socio-demographic attributes of both annotators and targets influence bias detection and labelling patterns across extensive datasets (Giorgi et al., 2025). This approach demonstrates the broader applicability of LLMs as analytical tools for identifying systematic biases in text classification and annotation workflows. Additionally, researchers have utilised multiple LLMs concurrently to audit content moderation systems, revealing how different models produce remarkably different classification outcomes for identical content (Fasching and Lelkes, 2025). These comparative analyses highlight inconsistencies across moderation systems, with variations particularly pronounced for specific demographic groups, underscoring the value of LLMs as tools for stress-testing and auditing bias in other AI systems deployed for content moderation and textual analysis.

In the domain of media bias, the research by Lin et al. (2024) critically scrutinises the trustworthiness of LLMs when deployed for bias prediction. This study adopts a dual analytical perspective, investigating both the inherent biases within the LLM systems themselves and concurrently evaluating their efficacy as detection tools for external media content. Methodologically, the authors employ a comparative framework, directly benchmarking LLM performance against human perception in core tasks such as political bias prediction and text continuation to identify nuanced variations in bias expression across topics. The principal finding reveals a significant disparity between LLM-generated bias assessments and human judgement, concluding that intrinsic biases inherited from the LLM architecture substantially impair objective detection capabilities. To address this, the study proposes practical strategies to reduce model bias, including fine-tuning and refining how models interpret and respond to input, with the goal of fostering more robust and equitable AI systems capable of mitigating the influence of misinformation.

The research presented by Elbouanani et al. (2025) explicitly positions LLMs as diagnostic tools for the detection of bias and opinion within text. This study, undertaken as participation in the CheckThat! 2025 evaluation campaign applied a focused benchmarking methodology to the task of multilingual subjectivity detection. The scope of the work involved the deployment of LLMs using few-shot learning and detailed prompt engineering strategies, including innovative approaches such as 'debate prompting', to evaluate their robustness against traditional fine-tuned Smaller Language Models (SLMs). A critical finding demonstrated that LLMs, when supported by carefully crafted few-shot prompts, could successfully match or even outperform SLMs, particularly proving resilient and effective when handling low-quality or inconsistently annotated multilingual data, such as the Arabic dataset used in the competition. Crucially, the authors noted that while advanced prompt engineering was explored, it often yielded limited benefits beyond a well-designed standard few-shot approach, yet their LLM-based system still secured top rankings across multiple language tracks, underscoring the potential of LLMs as highly adaptable instruments for complex text analysis tasks.

The study by Kumar and Singh (2024) employed a technical methodology where LLMs functioned as tools for bias mitigation within standard text classification tasks. The authors integrated LLMs with conditional Generative Adversarial Networks (cGANs), leveraging the advanced generative capabilities of the combined architecture to address fairness issues in classification models. The LLMs' role was crucial in enabling the cGAN to synthesise linguistically diverse and intentionally de-biased training examples, thereby augmenting the initial biased dataset. This approach aimed to rectify data deficiencies such as stereotypic associations or class imbalance prior to model training. The main findings validate this novel hybrid framework, demonstrating its effectiveness in enhancing the robustness of text classification models and achieving measurable improvements in key fairness metrics.

2.5 LLMs for Automated Text Annotation

The use of LLMs for automating text annotation has emerged as a prominent research area within NLP, primarily driven by the labour-intensive and costly nature of manual annotation. Whilst the potential applications are considerable, the literature reveals considerable variation in reported efficacy, methodological rigour and practical applicability across different annotation contexts. This section critically synthesises recent research, examining the performance claims and methodological limitations.

2.5.1 LLMs Annotating Capabilities

The emergence of LLMs has transformed data annotation, one of the most critical yet resource-intensive components of ML development. A recent survey by Tan et al. (2024) on LLM annotation suggest that models such as GPT-4 can achieve performance levels approaching or exceeding human annotators across various tasks. Unlike traditional crowdsourced annotation approaches, LLMs offer unprecedented capabilities for efficient, scalable and cost-effective text labelling. The advantages of LLM-based annotation systems include:

2.5.1.1 Remarkable Speed and Scalability

LLMs excel at processing vast quantities of text with exceptional speed (Csanády et al., 2024; Tan et al., 2024). Huang, T.-H. et al. (2024) argues that LLMs are an efficient way to annotate data and have the potential to partially or fully replace expensive human crowd workers, making LLM annotation highly scalable for large-scale NLP projects. The efficiency gains achieved by distilling the LLM logic into smaller, task-specific models translate directly to practical advantages, enabling the labelling of entire corpora in feasible timeframes (Csanády et al., 2024; Huang, T.-H. et al., 2024). Additionally, when LLMs are incorporated into active learning loops, they can intelligently determine which samples to annotate, further optimising the annotation process and reducing redundant labelling efforts (Zhang et al., 2023).

2.5.1.2 Few-shot Learning Excellence

LLMs demonstrate remarkable proficiency in few-shot and zero-shot learning scenarios, which is particularly valuable for annotation tasks. LLMs exhibit proficiency across a wide range of NLP tasks through minimal prompt-tuning. For instance, in specialised classification tasks on the United Medical Language System (UMLS) biomedical domain, Llama-2-70b-chat achieved 94.6% accuracy using just 1-shot prompting, demonstrating the model's ability to grasp complex task requirements with minimal examples (Csanády et al., 2024). This few-shot capability enables rapid task adaptation without extensive retraining, making LLMs particularly valuable for emerging or niche domains where labelled data is scarce (Tan et al., 2024).

2.5.1.3 High Annotation Accuracy and Performance

Multiple empirical studies confirm that LLMs produce high-quality annotations that can surpass or perform on par with human annotators (Csanády et al., 2024; Huang, T.-H. et al., 2024). On the IMDb sentiment classification task, Llama-2-70b-chat achieved 95.39% accuracy in zero-shot settings, with the best combined strategies reaching a state-of-the-art accuracy of 96.68% using fine-tuned RoBERTa-large (Csanády et al., 2024). Huang, T.-H. et al. (2024) demonstrate that LLM-based annotation improves performance on five out of eight datasets, showing an average enhancement of 12.9% compared to traditional annotation methods. Moreover, Zhang et al. (2023) show that task-specific models trained from LLM-generated labels can outperform the teacher LLM within only hundreds of annotated examples, achieving significantly more cost-effective annotation than traditional approaches.

2.5.2 LLMs Annotating Limitations

Pavlovic and Poesio (2024) conducted a comparative overview of twelve studies investigating LLM annotation capabilities, noting that whilst these models demonstrated "promising cost and time-saving benefits", the studies themselves suffered from considerable limitations, such as representativeness, bias, sensitivity to prompt variations and English language preference. This observation is particularly significant, as it suggests that positive findings in the literature may not generalise beyond the specific experimental conditions and datasets examined. Despite the strength of LLMs and cost effectiveness, there are several limitations, including:

2.5.2.1 Model Scale Does Not Ensure Annotation Reliability

The work by Haq et al. (2025) provides crucial evidence regarding the instability and reliability of using LLMs for text annotation, even in relatively structured classification scenarios such as NER. Their comparative analysis, which assessed models ranging from approximately 7 billion to 70 billion parameters across multiple datasets, revealed that annotation performance was not simply correlated with model size. Instead, it varied substantially based on the specific LLM architecture and the choice of embedding model. This finding directly contradicts the "deterministic claims" often propagated in broader

LLM survey literature, which may imply that increasing model size automatically translates to robust annotation capability. Crucially, the authors conclude that much of the success reported in the field may be due to current studies tending to "overrepresent performance on relatively tractable annotation tasks". This underscores a significant concern regarding the generalisability of positive results and highlights the necessity to direct research efforts "towards more challenging datasets" to obtain a realistic and critically informed assessment of LLMs' efficacy as reliable data annotators.

2.5.2.2 Inconsistent Performance Across Different Task Types

A critical gap in the literature concerns task-specific performance variation. Pangakis and Wolken (2024a) evaluated GPT-4 across 27 annotation tasks derived from 11 datasets in computational social science research, reporting "significant variation in LLM performance across tasks, even within datasets". This finding directly contradicts the homogenising tendency of survey literature, which often implies relatively consistent performance. More problematically, the authors found that "automated annotations significantly diverge from human judgment in numerous scenarios, despite various optimisation strategies such as prompt tuning", suggesting that technical interventions may have limited efficacy in addressing fundamental limitations.

2.5.2.3 Limitations in Cost Efficiency

Whilst cost reduction is frequently cited as a primary advantage (He et al., 2023; Zhang et al., 2023; Csanády et al., 2024; Huang, T.-H. et al., 2024; Kim, H. et al., 2024; Pangakis and Wolken, 2024b) the evidence base is surprisingly limited. The necessity of querying LLMs APIs for every data point makes direct annotation an unsustainable expense at scale. For instance, labelling 7,569 data points with GPT-4 can cost over \$1,200 ≈ £889.55, while using Llama-2-70b-chat to annotate millions of records could require nearly 367 days of computational time, demanding significantly greater computing power (Csanády et al., 2024). This high cost and lack of extensibility have driven the development of more efficient methods.

2.5.3 Human-in-the-Loop Integration in LLM Annotation

A key challenge is establishing what constitutes 'ground truth' in bias detection determining which instances genuinely represent bias when different algorithms may identify different patterns, as demonstrated by the varying accuracy rates across ML models in Kausar et al. (2023) study. Additionally, automated approaches that rely on linguistic markers may overlook subtle, context-dependent forms of bias that manifest through narrative framing, knowledge hierarchies or implicit assumptions rather than explicit language patterns.

Hitti et al. (2019) proposed developing standardised taxonomies and annotation frameworks for bias, recognising that consistency in bias identification is essential for advancing the field and enabling meaningful comparisons across studies. However, Albuquerque (2023) findings that bias perception varies significantly based on annotators' ethnic

backgrounds highlights the tension between standardisation and the inherently subjective, culturally situated nature of bias recognition. This underscores the continued necessity of human annotation and judgement in automated bias detection. The integration of human judgement with computational power appears particularly promising, as it combines the contextual sensitivity and cultural awareness of human review with the scale, consistency, and pattern-recognition capabilities of automated analysis. This hybrid approach allows researchers to leverage the efficiency of computational methods whilst maintaining the nuanced understanding that human reviewers bring to interpreting whether linguistic patterns constitute meaningful bias within specific educational and cultural contexts.

An important direction in the recent literature involves moving away from full LLM automation towards hybrid human-LLM systems. Pangakis and Wolken (2024a) emphasise that "human oversight is essential to ensure fair and reliable assessment outcomes", whilst their findings revealed that LLM suggestions did not accelerate annotation but did increase annotator confidence which raises questions about whether confidence constitutes a meaningful efficiency gain.

The prevalence of human-in-the-loop recommendations in the literature for example, Pavlovic and Poesio (2024) work may indicate that pure automation approaches have proven insufficiently reliable in practice, contradicting the more optimistic framing of LLM capabilities.

Despite advocacy for hybrid approaches, the literature provides limited direct comparisons of their effectiveness relative to either pure human annotation or pure LLM annotation. Existing studies typically evaluate specific hybrid systems such as collaborative annotation interfaces (Park, J. et al., 2024) without systematic comparison to baseline conditions, making it difficult to quantify the actual efficiency gains of human-LLM collaboration relative to established annotation workflows.

Regardless of the limited available evidence or the incorporation of human-in-the-loop strategies that appear to improve annotation outcomes, several persistent limitations continue to constrain the generalisability of findings on LLM-based annotation. A recurring set of challenges arises from issues of dataset representativeness, publication bias and inconsistent evaluation practices. Most existing research relies on relatively well-defined annotation tasks using standard benchmark datasets, as Ul Haq et al. (2025) highlight in their call for work on more challenging, domain-specific and ambiguous datasets that better reflect real-world complexity. However, the predominance of such tractable problems means that reported results may overstate LLMs' true annotation efficacy. This bias is compounded by selective reporting and publication bias, with studies more likely to emphasise positive outcomes or novel applications of LLMs, while negative or mixed findings such as those noted in systematic reviews by Emirtekin (2025), which stress the need for human oversight and context-dependent effectiveness are less frequently published. Furthermore, methodological inconsistency in evaluation metrics (e.g., Quadratic Weighted Kappa (QWK), Intraclass Correlation Coefficient (ICC), Pearson correlation) further obscures comparability, while some studies report near-human agreement (QWK

= 0.99, r = 0.95), others show much lower reliability (ICC = 0.45, r = 0.38), yet few investigate the sources of this variability (Emirtekin, 2025). Collectively, these issues narrow dataset scope, selective reporting and heterogeneous evaluation limit the interpretability and transferability of current evidence regarding LLMs' annotation capabilities.

2.6 Integrating RAG and LLMs in Learning Systems

RAG offers a compelling paradigm for integrating LLMs into structured learning environments by coupling their generative capacity with retrieval over domain-specific, institutionally curated content. Rather than relying solely on the LLM's internal parameters, RAG first retrieves relevant passages from a knowledge base (e.g., course materials, textbooks, knowledge graphs) and then conditions generation on those retrieved contexts.

Empirical work underscores the benefits of this integration. For instance, Cohn et al. (2025) introduce Log-Contextualised RAG (LC-RAG) in a pedagogical agent for STEM learning, demonstrating improved retrieval and personalised, curriculum-aligned guidance. Kuratomi et al. (2025) build a RAG-based assistant for a university, showing that providing correct document chunks to the LLM significantly improves response accuracy. Furthermore, systematic reviews indicate that RAG is widely being adopted in educational applications, because it "improves factual accuracy and enables dynamic knowledge updates," while also identifying challenges like hallucination, retrieval accuracy and resource alignment (Li et al., 2025).

Alsafari et al. (2024) demonstrate that LLM-powered teaching assistants, when coupled with a retrieval layer, can interpret a student's intent, fetch relevant lecture slides, textbook excerpts or lab manuals and then produce concise, context-aware explanations that align with the course's pedagogical goals. Similarly, Abbas and Atwell (2025) show that a RAG-enabled feedback system can retrieve rubric criteria, exemplar solutions and prior grading data to generate detailed, criterion-based assessment comments, thereby increasing the consistency and transparency of automated feedback for large cohorts.

Collectively, this body of research demonstrates that RAG-augmented LLMs in educational contexts yield both theoretical validity and practical efficacy, producing contextually accurate, pedagogically sound outputs grounded in authoritative and current domain knowledge.

2.7 Conclusion

This chapter provided an overview of the key research areas underpinning this thesis. It explored the foundational concepts and taxonomies of bias in NLP, examined how ethnic and gender biases manifest in educational texts and outlined the technical landscape of using LLMs for bias investigation and automated text annotation, including their capabilities, limitations, the integration of human-in-the-loop approaches and the integration of

RAG with LLMs. Subsequent chapters provide domain-specific background and methodological details pertinent to their respective investigations.

Chapter 3

3 Bias in Higher Education Online Content: Datasets and Annotations

3.1 Introduction

3.1.1 The Central Hypothesis: Higher Education Texts Are Not Neutral

Higher education institutions present themselves as bastions of objective knowledge, where facts are separated from opinion and scholarly discourse maintains rigorous neutrality (Hartley, M. and Morphew, 2008). This assumption of objectivity permeates academic culture, from lecture halls to published research, creating an expectation that higher educational texts represent unbiased, authoritative sources of knowledge. However, this perceived neutrality may be more myth than reality.

The central hypothesis driving this research challenges a fundamental assumption about academic discourse: higher education texts are not neutral. Despite claims of objectivity, educational materials carry embedded biases that shape how knowledge is presented, whose voices are amplified, and which perspectives are marginalised (Rathbun and Turner, 2012). These biases operate across multiple levels of educational discourse, from institutional communications to curriculum design to direct learning materials.

Academic institutions, like all organisations, are shaped by power structures that determine which voices receive prominence and how information is framed (Ahmed, S., 2012). The myth of neutrality obscures these dynamics, creating a false sense of objectivity that can perpetuate existing inequalities rather than challenge them. As Schucan Bird and Pitman (2020) demonstrate in their analysis of UK universities reading lists, even seemingly neutral academic selections reflect systematic patterns of exclusion and inclusion that favour particular demographic groups and perspectives.

The hidden influence of institutional voice manifests in multiple ways throughout the educational ecosystem. University news articles, whilst appearing to provide factual reporting on institutional activities, may systematically favour certain groups in terms of visibility and positive framing (Wagner et al., 2015). Curricular choices, from reading list compositions to module content, reflect implicit value judgements about whose scholarship deserves attention and which perspectives merit inclusion (Bhambra et al., 2018). Pedagogical framing shapes how students encounter and interpret knowledge, potentially reinforcing existing social hierarchies through seemingly objective educational content (Mahmud and Gagnon, 2023).

3.1.2 Data Framework for Bias Analysis in Higher Education Online Content

This research employs a three-stage approach to systematically uncover bias in higher education online contents, progressing from broad institutional patterns to specific learning interactions. The investigation follows three distinct but interconnected domains of online educational content. First, institutional representation examines how universities present themselves and their community members through official university news articles and institutional communications. This level captures the institutional voice and reveals patterns of visibility and prominence across different demographic groups. Second, curricular diversity investigates the composition of academic reading lists and syllabi, examining whose scholarship is valued and which perspectives are privileged in formal curricula. This level exposes the gatekeeping mechanisms that determine which knowledge is considered legitimate and worthy of academic study. Third, direct learning materials analysis focuses on the actual educational content that students engage with, including online educational resources, and other instructional materials, to identify embedded biases that may directly influence learning. This category also includes transcripts of recorded university lectures obtained from a UK higher education institution.

This progression from institutional voice to curricular choices to direct pedagogical content ensures comprehensive coverage of the online educational ecosystem. Each level operates with different constraints and stakeholders, yet all contribute to shaping student perceptions and understanding (Apple, 2014). The systematic examination of these three domains provides a complete picture of how bias operates across the entire online higher educational pipeline.

The movement from institutional representation through curricular diversity to direct learning materials represents an increasing proximity to student learning experiences. Institutional representation, whilst important for public perception and community identity, operates at a remove from daily educational interactions. University news articles serve dual functions as internal communications for staff and students whilst also influencing external perceptions of the institution (Hartley, M. and Morphew, 2008). These official institutional communications shape broader narratives about academic achievement and institutional priorities, primarily influencing external perceptions rather than direct learning outcomes.

Curricular diversity moves closer to the educational core by examining the knowledge sources that educators select for student engagement. Reading lists and syllabi represent deliberate choices about which authors, perspectives and intellectual traditions students will encounter (Schucan Bird and Pitman, 2020). These selections directly influence the range of viewpoints and methodological approaches that students consider legitimate within their field of study. The demographic composition of reading lists has been shown to correlate with student perceptions of who can be successful in academic careers (Phull et al., 2019). When students consistently encounter scholarship produced predominantly

by one demographic group, it reinforces implicit assumptions about whose knowledge is valued and who belongs in academia, which can discourage students from underrepresented backgrounds from envisioning themselves in scholarly roles.

Direct learning materials represent the closest point of contact between bias and student learning. Educational texts, case studies and instructional content shape not only what students learn but how they understand concepts and their applications (Sleeter and Grant, 2008). At this level, bias can influence student perceptions through representation patterns, stereotypical portrayals and evaluative language that frames certain groups or perspectives as more or less valuable (Banks, 2015).

3.1.3 Objectives of the Chapter

This chapter aims to achieve three main objectives in the study of bias in higher education online content. First, it seeks to systematically define the key types of bias within higher education discourse and outline targeted detection methodologies for each. Second, it aims to build a comprehensive investigative framework capable of capturing these biases across diverse educational resources, including news articles, reading lists and online course content. Third, it sets out to design annotation frameworks that integrate human expertise and AI-assisted methods, establishing the basis for scalable and reproducible bias identification.

3.1.4 Contributions of the Chapter

This chapter makes three distinct contributions to the research on educational bias. First, it delivers a structured conceptual framework for understanding and categorising bias in higher education contexts. Second, it introduces newly constructed datasets that span university news articles, reading lists from multiple cultural contexts and online course materials across humanities and STEM domains, providing a solid foundation for empirical bias analysis. Third, it presents a hybrid annotation methodology combining manual and automated approaches resulting in gold standard datasets with consistent labelling protocols that support future large-scale bias detection and research replication.

Therefore, this chapter aims to achieve the main contributions introduced in Chapter 1, namely:

1. Establishing a bias classification framework for online higher education resources (see Section 1.3.1).
2. Presenting three novel datasets (see Section 1.3.2).
3. Developing LLM-based bias annotation strategies (see Section 1.3.3.1).
4. Proposing a hybrid human-LLM annotation methodology (see Section 1.3.4.1).

This chapter integrates multiple contributions to establish the foundational infrastructure, including bias categorisation frameworks, datasets and annotations, upon which subsequent chapters build through specialised investigations of individual datasets.

3.2 Types of Educational Bias

Educational bias manifests in multiple interconnected forms, each contributing to the perpetuation of inequalities within higher education. This research focuses on three critical types of bias that systematically undermine the assumed neutrality of academic discourse: representation bias, stereotype bias and linguistic bias.

These three categories are utilised to classify gender and ethnic biases in different online higher educational texts. These categories are well recognised in the literature for their ability to capture both overt and subtle forms of bias, which is crucial when analysing the complex and nuanced characteristic of higher education content. Table 1.1 in Chapter 1 provides illustrative examples of these bias category manifestations in educational contexts.

3.2.1 Representation Bias

Representation addresses the visibility and portrayal of different gender and ethnic groups within academic content. More definitions from the literature:

- "Representation bias in data can happen due to various reasons ranging from historical discrimination to selection and sampling biases in the data acquisition and preparation methods" - (Shahbazi et al., 2023).

- "Racial, ethnic and gender representation in an academic setting means that teachers, professors, and other leaders reflect the demographics of the student body in the educational and professional spaces that they serve" - (Ijoma et al., 2022).

This form of bias operates through the systematic inclusion and exclusion of voices, perspectives, and demographics within educational materials, functioning as an invisible gatekeeper that determines who gets to speak and whose knowledge is legitimised within academic discourse (Schucan Bird and Pitman, 2020). Numerous studies have documented the underrepresentation of women and ethnic minorities, particularly in leadership roles or STEM-related fields (Gooden and Gooden, 2001; Lee, J.F. and Collins, 2008). This underrepresentation perpetuates harmful stereotypes about the societal roles of these groups. Chisholm (2018) emphasises the importance of inclusive representation in educational materials, advocating for a more balanced approach that recognises the contributions and experiences of all groups, especially those that have historically been marginalised.

Another example of representation bias in higher education, EU research funding evaluations have been shown to penalise proposals that include a higher proportion of female Principal Investigators, resulting in lower reviewer scores and reduced funding success for gender-balanced teams. This implicit bias effectively makes gender diversity a liability rather than an asset in the allocation of EU research grants (Standing Working Group on Gender in Research and Innovation, 2018).

3.2.1.1 Whose Voices are Represented, Amplified or Marginalised in Discourse?

Recent analysis of university textbooks across multiple countries shows that female representation averages only 40.4% in combined textual and pictorial indicators, with even lower representation in prestigious professional roles (Islam and Asadullah, 2018). This underrepresentation is particularly pronounced in higher education contexts, where Bachore and Semela (2022) found that gender bias persists from primary education through to university-level materials, creating what researchers term the "hidden curriculum" of higher education.

3.2.1.2 How Does the Transition from Author Demographics to Perspective Inclusion Affect Representation?

Representation bias extends beyond simple demographic counting to encompass the perspectives and epistemologies that are privileged within academic discourse. Jehle et al. (2024) demonstrates that across European educational contexts, textbooks consistently underrepresent not only female authors but also alternative viewpoints that challenge traditional power structures. This creates what Bhambra et al. (2018) describe as "epistemic injustice" where certain forms of knowledge are systematically excluded from academic legitimacy.

3.2.1.3 Why Representation Shapes What Students See as Possible?

The psychological impact of representation bias is profound. Students develop their academic and professional aspirations based partly on the models of success they encounter in educational materials (Zarrett and Lerner, 2008). When certain groups are consistently underrepresented or portrayed in limited roles, this constrains students' perceptions of what is achievable for people like themselves. This phenomenon is particularly damaging in STEM fields, where Leslie et al. (2015) found that expectations of brilliance often associated with male representation significantly influence gender distributions across academic disciplines.

3.2.2 Stereotype Bias

Stereotype refers to the attribution of fixed, oversimplified traits to specific gender or ethnic groups. In higher education materials, this often manifests in how individuals, groups or historical figures are depicted. More definitions from the literature:

- "A gender stereotype is a generalised view or preconception about attributes or characteristics, or the roles that are or ought to be possessed by, or performed by, women and men. A gender stereotype is harmful when it limits women's and men's capacity to develop their personal abilities, pursue their professional careers and/or make choices about their lives" - (OHCHR, no date).

- "Stereotyped language is any that assumes a stereotype about a group of people" - (Purdue University OWL no date).

Stereotype bias operates through the embedding of these fixed assumptions within seemingly neutral educational descriptions, working through subtle linguistic choices that reinforce existing social hierarchies whilst maintaining the appearance of objectivity (Hinton, 2017).

3.2.2.1 In What Ways are Fixed Assumptions Embedded Within Claims of Neutrality?

Educational materials frequently contain stereotypical portrayals disguised as factual descriptions. For example, textbooks have been shown to reinforce traditional gender roles by portraying women in nurturing or passive capacities, whilst men are often depicted as leaders or intellectual authorities (Crabb and Bielawski, 1994; Lee, J.F. and Collins, 2008). Similarly, ethnic minorities are frequently represented through colonial or subservient lenses, which restricts their representation (Schoeman, 2009). This issue persists in contemporary educational materials, as highlighted by Jackie and Lee (2011), who demonstrate that racial and gender stereotypes continue to affect the portrayal of both genders and ethnic groups in modern textbooks. These patterns suggest that stereotyping bias operates below the threshold of conscious awareness amongst content creators.

3.2.2.2 How Educational Content Perpetuates Societal Stereotypes?

The perpetuation of stereotypes through educational content occurs through what researchers' term 'linguistic intergroup bias' (Maass, 1999). This involves systematic differences in how behaviours and characteristics are described depending on group membership. Educational materials often describe male achievements using concrete, stable language whilst female accomplishments are portrayed using more qualified or temporary framing, thereby reinforcing stereotypical expectations about gender roles and capabilities.

3.2.2.3 How Can the Use of Language Reinforce Limitations on Ambition and Potential?

Stereotyping in educational content has measurable impacts on student outcomes. Experimental research by Copur-Gencturk et al. (2022) demonstrates that teachers' stereotypical biases, reinforced through educational materials, directly influence student performance and self-concept. When educational content consistently portrays certain groups in stereotypical roles, it creates what researchers call 'stereotype threat', limiting students' academic engagement and career aspirations.

3.2.3 Linguistic Bias

Linguistic bias represents perhaps the most subtle form of educational bias, operating through evaluative framing and emotional undertones that influence perception whilst maintaining the appearance of neutrality. More definitions from the literature:

- According to the Oxford research encyclopaedia of communication a linguistic bias is defined as "A systematic asymmetry in word choice that reflects social-category cognitions applied to a group or person." - (Beukeboom and Burgers, 2017).

- "Linguistic bias refers to the preference for or prejudice against specific languages, dialects, or features of language use. Language varieties that are commonly denigrated … as well as features associated with racial or ethnic groups …" - (Shin et al., no date)

The language used in educational texts can perpetuate bias, often through the use of gendered pronouns or male-default language, reinforcing a patriarchal view of society (Gupta and Yin, 1990; Hsu, 1992). Additionally, the tone employed to describe ethnic minorities can be patronising or dismissive, contributing to their marginalisation (Schoeman, 2009). Studies examining language patterns in higher education materials reveal that these texts frequently fail to challenge ingrained discriminatory language practices (Schoeman, 2009; Jackie and Lee, 2011). A recent study by Huang, P. and Liu (2024) calls for a shift towards more equitable language use in education, emphasising that linguistic choices significantly impact how students perceive societal roles and identity groups.

3.2.3.1 In What Ways Can Factual Reporting Conceal Subjective or Evaluative Framing?

Educational materials often employ evaluative language disguised as objective description. This manifests through word choices, sentence structures and framing devices that subtly favour certain perspectives whilst marginalising others. Research on academic language reveals systematic patterns of bias in how different groups and their achievements are described, with dominant groups receiving more positive and concrete language than marginalised groups (Quinn, 2020).

3.2.3.2 In What Ways Can Emotional Undertones Affect the Interpretation of Information?

The emotional framing of educational content significantly influences student perceptions and attitudes. Fan, Y. et al. (2019) demonstrate that subtle variations in tone and language choice in educational materials can activate implicit biases, affecting how students

perceive different groups and their contributions. This is particularly problematic because such biases operate below the threshold of conscious awareness, making them difficult to identify and challenge.

3.2.3.3 How Institutions Use Language to Construct Preferred Narratives?

Educational institutions employ language strategically to construct particular narratives about knowledge, authority and social order. This occurs through the selection of voices, the framing of historical events and the presentation of contemporary issues. Research on institutional discourse reveals that universities often use apparently neutral language to maintain existing power structures whilst appearing to embrace diversity and inclusion (Ahmed, S., 2012).

3.2.4 The Interconnected Nature of Educational Bias

Research on intersectionality demonstrates that multiple forms of bias create cumulative effects that exceed the simple addition of individual biases (Crenshaw, K., 2013). As Crenshaw explains, intersectionality operates as "a lens, a prism, for seeing the way in which various forms of inequality often operate together and exacerbate each other" (Merriam-Webster, 2021). Studies on intersectional implicit bias show that biases compound across multiple categories asymmetrically, with effects that are "not necessarily additive, or even multiplicative, but rather, interactive in complex ways" (Connor et al., 2023).

In educational contexts, these interconnected biases create what researchers term "intersectional invisibility" (Purdie-Vaughns and Eibach, 2008). When women are underrepresented in educational materials (representation bias), portrayed in stereotypical roles when they do appear (stereotype bias) and described using less authoritative language (linguistic bias), the combined effect systematically marginalises their contributions to knowledge.

Traditional approaches to bias detection have focused on single dimensions, typically demographic representation, but research demonstrates significant limitations in such approaches (Baker and Hawn, 2021). Single-dimension analyses often miss "the subtle but pervasive ways in which bias operates through language and framing" and fail to capture how multiple bias types interact (Garg et al., 2018). By examining all three types simultaneously, this study addresses the recognised need for multi-dimensional approaches to bias detection in educational materials. The interconnections among bias categories are exemplified in Table 1.1 (Chapter 1).

Detecting these interconnected forms of bias presents significant methodological challenges. Traditional content analysis approaches rely on surface-level indicators and "miss the subtle linguistic and contextual cues through which bias operates" (Blodgett et al., 2020). Educational bias often operates through omission rather than commission; what is not said or included can be as significant as what is explicitly stated. This study therefore draws on a diverse range of data sources to capture a more complete picture of how these biases manifest. Each subsequent chapter examines one or more of the identified bias types across different educational contexts, using varied sources to build a comprehensive

understanding of how each bias type contributes to non-neutrality within higher education discourse.

3.3 Datasets: Three Sources for Investigating Bias in Higher Education Online Content

3.3.1 University News Articles: Institutional Perspective

3.3.1.1 Why University News?

Universities function as significant cultural institutions that shape public discourse through their official communications (Hartley, M. and Morphew, 2008). University news articles serve as a critical window into institutional identity, representing how higher education institutions present themselves to both internal and external audiences. These articles constitute the official narrative of academic achievement, research breakthroughs, and institutional priorities that universities wish to project.

Unlike traditional media outlets that operate under commercial pressures and editorial independence, university news is typically managed by internal communications teams that prioritise institutional branding and stakeholder interests (Zerfass et al., 2017). This creates a unique discourse environment where bias may operate through selection decisions about whose achievements are highlighted, how different groups are portrayed and what language is used to frame institutional success. Educational institutions are often thought of as progressive spaces; however, the institutions themselves operate in a societal and political context that reflect beliefs wrong or right that societies hold (Smith, J. et al., 2023).

Research indicates that institutional bias operates through procedures and practices of particular institutions that result in certain social groups being advantaged or favoured and others being disadvantaged or devalued (Oxford, 2025). This need not result from conscious prejudice but rather from following existing norms and practices. University news articles therefore provide an ideal dataset for investigating how institutional communication may inadvertently perpetuate representational inequalities.

3.3.1.2 Dataset Overview

A corpus of University of Leeds News Articles (UoL-NA) consisting of 5,768 articles collected from the university official news website, spanning April 2009 to March 2023. This scale was deliberately chosen to balance robust statistical power for detecting subtle bias patterns across demographic and temporal subgroups, while remaining computationally manageable for comprehensive NLP and statistical analyses. Such substantial sample sizes are critical in higher education bias research to establish reliable patterns and avoid pitfalls like multiple comparison errors (Fan, Y. et al., 2019).

UoL-NA covers diverse topics including academic achievements, administrative announcements, research breakthroughs and campus events. The 14-year span captures institutional communication shifts amid social movements for diversity, inclusion and external factors such as policy changes and the COVID-19 pandemic, allowing longitudinal bias analysis (Llorens et al., 2021).

Data were collected via automated web scraping with formal permission from the university's communications administration, ensuring ethical compliance and data quality. Preprocessing steps such as lowercasing, removing punctuation, stop words, duplicates and categorising articles prepared the dataset for bias detection, supporting detailed subgroup and temporal analyses.

3.3.1.3 Data Size

The comprehensive scale of the UoL-NA was driven by two primary methodological requirements. Firstly, the longitudinal analysis spanning April 2009 to March 2023 necessitated sufficient data points across temporal periods to detect meaningful patterns in institutional communication evolution. Given that the investigation examined how bias patterns shifted in response to social movements, policy changes and external factors such as the COVID-19 pandemic, a robust dataset was essential to capture these temporal variations whilst maintaining statistical validity across different time periods.

Secondly, the gender distribution analysis required substantial sample sizes to ensure reliable detection of subtle representational patterns. Statistical analysis of gender representation across different contexts (academic achievements, administrative announcements, research coverage) demands large datasets to distinguish genuine bias patterns from random variation (Cohen, 2013; Field, Andy, 2024). With female representation in academic contexts often constituting a minority proportion, particularly in senior leadership roles and STEM fields (Moss-Racusin et al., 2012; Sheltzer and Smith, 2014), detecting statistically significant disparities requires sufficient cases to achieve adequate statistical power (Button et al., 2013). Research on gender representation in academic media demonstrates that meaningful pattern detection requires substantial sample sizes due to the intersectional nature of bias manifestations across different institutional contexts (Wagner et al., 2015; Llorens et al., 2021). The 14-year span with nearly 6,000 articles provides the necessary scale for robust testing and temporal trend analysis whilst enabling meaningful subgroup analysis across different types of institutional communication (Kim, D., 2013; McHugh, 2013)

3.3.1.4 What UoL-NA Reveals About Bias?

The primary bias category examined in this data is gender bias, alongside the following secondary bias categories.

- **Representation: Who Gets Featured in University News and How?**

The primary bias investigation focuses on representational patterns within university news coverage. Gender bias has a substantial negative impact on the careers, work-life balance and mental health of underrepresented groups in science, manifesting through various institutional practices including communication and recognition (Carnes et al., 2012).

Our analysis examines which individuals and groups receive coverage in university news, with particular attention to gender representation patterns. This includes investigating the frequency with which male and female academics, students and staff are featured, the contexts in which they appear and the types of achievements or activities that generate coverage for different demographic groups.

The representational analysis extends beyond simple demographic counting to examine the nature of representation. Research indicates that student evaluations and institutional recognition systems can reflect biases that affect how different groups are perceived and valued within higher education contexts (Peterson et al., 2019). Our investigation explores whether certain types of achievements or roles are more likely to generate news coverage for particular gender groups, potentially revealing institutional priorities and values.

- **Linguistic Bias: Subjectivity in Supposedly Objective Institutional Reporting**

The second bias investigation examines the language employed in institutional communication. While university news articles are typically perceived as objective reporting of facts and achievements, research suggests that positivity bias is the tendency, in some forms of published higher education research, to only or chiefly report examples of initiatives or innovations that worked and received positive evaluations (Tight, 2023).

Our analysis investigates subjectivity and evaluative language within institutional reporting. This includes examining whether certain groups or achievements are described using more positive, authoritative, or detailed language and whether subtle linguistic choices may influence reader perceptions of different individuals or groups.

Bias is any tendency which prevents unprejudiced consideration of a question, occurring when systematic error is introduced into sampling or testing by selecting or encouraging one outcome or answer over others (Pannucci and Wilkins, 2010). In institutional communication, such bias may manifest through seemingly neutral language choices that nonetheless frame achievements, individuals, or groups in subtly different ways.

- **Stereotype Bias: Reinforcing Social and Gendered Norms in News Reporting**

The third bias investigation examines how university news articles may perpetuate or reinforce stereotypical representations of individuals and groups. Stereotype bias occurs when reporting reproduces social assumptions about gender, ethnicity, age or professional

roles, leading to the reinforcement of dominant cultural narratives rather than presenting an impartial or balanced view (Entman & Rojecki, 2001). In university or organisational contexts, such bias can shape perceptions of who is seen as a "typical" academic, leader or achiever.

Research in media studies has shown that news coverage often reflects and amplifies existing social hierarchies through language and framing (van Dijk, 1991; Gill, 2007). For instance, female academics are more frequently portrayed in relation to teaching, pastoral care or family responsibilities, while male academics are more often associated with leadership, research excellence or innovation (Lösch et al., 2022; North, 2016).

In the context of university reporting, stereotype bias may manifest through the repeated pairing of certain gender with particular achievements or roles. For example, featuring male staff members in stories about technological breakthroughs while women appear in community or wellbeing initiatives subtly reinforces gendered expectations. Visual and textual framing such as choice of imagery, quotations or descriptors can further accentuate these biases by presenting one group as authoritative and another as supportive or secondary (Macharia, 2020).

3.3.2 University Module Reading Lists: Curriculum Content Sources

Reading lists, whilst appearing as simple bibliographic compilations, may function as gatekeepers that influence which voices students encounter and which perspectives are legitimised within academic disciplines (Schucan Bird and Pitman, 2020). The "Why is My Curriculum White?" campaign highlighted how UK higher education consistently privileges European and male authors, marginalising diverse perspectives that could enrich student learning (Charles, 2019). Reading lists therefore represent a critical site for investigating representation bias in higher education.

3.3.2.1 Datasets Overview

The investigation reading lists examines three different reading lists from two different higher educational institutions: the University of Leeds, Leeds, UK and Imam Mohammad Ibn Saud Islamic University, Riyadh, Saudi Arabia. This selection was informed by my direct institutional engagement as an MSc and PhD student at the University of Leeds and as a Lecturer at Imam University, which provided both practical access to educational materials and contextual understanding of pedagogical frameworks within these institutions.

- **University of Leeds, United Kingdom - Arabic, Islamic and Middle Eastern Studies reading List (AIMESRL).**

AIMESRL dataset comprises 212 readings extracted from undergraduate modules at the University of Leeds. These readings span introductory and advanced courses covering Islamic history, Arabic literature, Middle Eastern politics and religious studies. The

materials were sourced from publicly available module handbooks and official reading lists published between 2020 and 2023.

- **Imam Mohammad Ibn Saud Islamic University, Saudi Arabia - Qur'an Sciences and Islamic Education Reading List (QSIERL).**

QSIERL contains 73 readings from modules taught at Imam Mohammad ibn Saud Islamic University. This dataset focuses specifically on traditional Islamic sciences including Quranic exegesis (tafsir), Islamic jurisprudence (fiqh) and Prophetic traditions (hadith). The readings were obtained from internal programme documentation and departmental websites covering the period 2021-2023.

- **University of Leeds, United Kingdom - Disability Studies Reading List (DSRL).**

The DSRL used across eight undergraduate modules at the University of Leeds. These modules span a range of disciplines, including law, literature, sociology and education. The reading lists extracted from the following modules:

- Disability Law
- Disability Rights and the International Policy Context
- Disability Studies: An Introduction
- Inclusion and Special Educational Needs and Disability
- Inclusive Education
- Extra Ordinary Bodies: Disability, Medicine and Normalcy in Contemporary Literatures
- Medical Humanities: Representing Illness, Disability and Care
- States of Mind: Disability, Cognitive Impairment and Mental Health in Contemporary Culture

The compiled reading lists contain a total of 679 items, reflecting a broad and interdisciplinary academic focus.

3.3.2.2 Rationale for Dataset Selection

- **Cross-Cultural Validation: Western vs. Middle Eastern Islamic Educational Content**

Comparing the University of Leeds and Imam University reading lists enables investigation of how different cultural and educational traditions construct knowledge. Research suggests that Western academic institutions often exhibit Eurocentric bias in their curriculum design (Bhambra et al., 2018), whilst Islamic universities may prioritise traditional scholarly authorities and Arabic-language sources (George, 1981; Saliba, 2007). This comparison allows examination of whether representation patterns vary across cultural

contexts or reflect universal academic hierarchies, building on research that suggests academic knowledge production remains influenced by both local traditions and global academic structures (Alatas, 2003; Connell, 2020).

- **Additional Diversity Data: Disability Studies Across Multiple Disciplines**

The DSRL serves as an additional dataset for investigating author diversity patterns within a single institutional context. Unlike the cross-cultural comparison between Leeds and Imam University, this dataset enables examination of representation across multiple academic disciplines within the same university system. As the 679 readings span eight different modules across various disciplines providing a broader sample for analysing whether author diversity patterns are consistent across different academic subjects or vary by disciplinary tradition.

- **Strategic Rationale for Cross-Cultural Domain Selection**

The selection of Islamic and Disability studies represents a deliberate methodological strategy designed to examine bias detection across fundamentally different epistemological and cultural frameworks. These domains were chosen not despite their specificity, but precisely because of their unique characteristics that illuminate broader patterns of educational bias.

- **Islamic Studies as A Critical Test Case for Cross-Cultural Bias Detection**

The inclusion of Islamic Studies serves multiple strategic purposes beyond simple cross-cultural comparison. Firstly, this field represents a domain where Western academic institutions have historically struggled with orientalist perspectives and cultural representation issues (Said, 1978). By examining reading lists in this area, the research can assess whether bias detection methodologies developed primarily within Western academic contexts can accurately identify subtle forms of cultural marginalisation and epistemic exclusion that may be invisible to culturally homogenous analytical frameworks.

Secondly, the comparison between University of Leeds and Imam Mohammad Ibn Saud Islamic University provides a unique opportunity to examine how different institutional contexts construct knowledge hierarchies. Western universities teaching Islamic studies may inadvertently privilege certain scholarly traditions or interpretative frameworks, whilst Islamic universities may exhibit different patterns of gender or sectarian representation. This comparison enables investigation of whether bias manifestations are universal features of academic discourse or culturally specific phenomena that require localised detection approaches.

- **Disability Studies as a Lens for Intersectional Bias Analysis**

The Disability studies reading lists offer a particularly valuable dataset for examining how academic disciplines that explicitly focus on marginalised communities construct their

own knowledge boundaries. This field provides insights into whether academic areas that position themselves as advocating for inclusion nonetheless reproduce subtle forms of bias in their scholarly selections. The interdisciplinary nature of disability studies spanning law, literature, sociology and education enable examination of how bias patterns vary across different academic traditions whilst maintaining thematic coherence.

- **Transferability to Broader Higher Education Contexts**

These seemingly specialised domains actually function as concentrated environments where broader educational bias patterns become more visible and measurable. Cultural studies programmes, religious studies departments and interdisciplinary fields focused on marginalised communities represent significant portions of contemporary higher education curricula. The methodological approaches developed through analysis of these specific reading lists provide templates for examining bias across other specialised academic areas including ethnic studies, gender studies, postcolonial literature and international development programmes.

Furthermore, the bias detection techniques validated through these culturally and epistemologically diverse datasets demonstrate robustness across different knowledge traditions, enhancing confidence in their applicability to mainstream academic disciplines. If bias detection methodologies prove effective in identifying subtle patterns within both Western orientalist discourse and traditional Islamic scholarship frameworks, they provide strong evidence for broader applicability across diverse educational contexts.

The findings from these specific domains contribute to understanding how academic gatekeeping operates across different cultural and institutional settings, revealing whether representation bias, stereotyping and evaluative language patterns represent universal features of educational discourse or culturally contingent phenomena requiring adapted analytical approaches.

3.3.2.3 What This Data Reveals About Bias?

The primary bias categories examined in this dataset are gender and ethnicity biases, followed by representation bias as a secondary category. The analysis of these reading lists focuses particularly on representation bias specifically, whose voices are included, amplified, or marginalised within university curricula. This is explored through two complementary lenses:

- **Reading List Representation – Whose Scholarship Is Valued?**

This perspective examines the demographics of authors whose work appears on reading lists, asking whose scholarship is prioritised and whose is overlooked. Previous research has documented significant gender and ethnic disparities in UK higher education reading lists. For example, Phull et al. (2019) found that International Relations curricula at British universities overwhelmingly featured white, male, Western authors, with minimal

inclusion of female scholars or non-Western perspectives. The present investigation builds on such findings to explore whether similar patterns emerge across different cultural contexts and academic disciplines.

- **Content Representation – Which Perspectives Are Privileged?**

Beyond author demographics, the thematic content of assigned readings is examined to determine which perspectives and narratives are foregrounded. Research indicates that even when reading lists include diverse authors, the selected works may still reinforce dominant viewpoints while sidelining critical or alternative perspectives (Ahmed, S., 2012). This study therefore considers both who is being read and what ideas are given prominence within these academic conversations.

3.3.2.4 Data Extraction Challenge and Manual Curation

The data was manually extracted from publicly available module handbooks and reading materials provided by each institution. For the AIMESRL and DSRL datasets, reading lists were retrieved from institutional repositories and official module descriptions, which included detailed bibliographic references for each assigned text. For QSIERL, the reading lists were sourced from internal programme documentation, departmental websites and digital copies of module outlines.

Once collected, the bibliographic entries for QSIERL and AIMESRL underwent a thorough cleaning and standardisation process to isolate author names, titles and publication metadata. This process involved removing duplicates and correcting formatting inconsistencies to ensure consistency across entries. A key methodological decision was made regarding multi-author works where a reading had multiple authors, the entry was excluded from the dataset. This decision was made to avoid duplication and ensure that each reading list item could be clearly attributed to a single author for the purposes of demographic analysis. Consequently, these datasets include only single-author readings. After cleaning and standardisation, each entry included the following fields: Type (e.g., book, journal article, chapter, etc.), Title, Author, Gender and Ethnicity. This structured format enabled consistent comparison and further analysis across both institutional datasets.

However, the DSRL dataset required a different approach due to the prevalence of edited volumes. Many entries in this dataset referenced chapters within edited volumes and in such cases, the chapter author was prioritised for demographic analysis rather than the book's overall author or editor. This approach better reflects the actual contributor of the assigned material. Unlike the other datasets, where entries with multiple authors were excluded, multi-author works in DSRL were retained within the same record. Importantly, this did not lead to duplication of reading list items; instead, multi-author works were counted once, with multiple demographic labels recorded where applicable. After cleaning and structuring, each DSRL entry included the following fields: Title, Chapter Title, Author, Chapter Author, Gender and Ethnicity.

Following the extraction of all the reading list entries (i.e., AIMESRL, QSIERL and DSRL), author-level gender and ethnicity demographic information was manually added wherever possible. For each listed author, efforts were made to identify gender and ethnicity using publicly available academic profiles, institutional biographies, publication records and other online sources. Where this information was not explicitly available or remained unclear, entries were marked as "unclear." Gender was classified using a three-category framework: male, female and non-binary, based on how the author self-identified or was described in institutional or biographical sources. This classification aimed to respect how individuals are represented publicly, whilst recognising the limitations of inferring gender from secondary information.

The approach to ethnicity classification varied between datasets. For AIMESRL and QSIERL, ethnicity was inferred based on contextual clues, including nationality, cultural identifiers and institutional affiliations. Six broad categories were developed: Middle Eastern, Asian, European, North American, African and unclear. Notably, authors from the Middle East were classified separately from Asian or African categories despite the geographical overlap. This decision reflects a region-specific cultural, linguistic and academic identity that is often treated as a distinct area of study, particularly within Arabic and Islamic Studies. Classifying Middle Eastern authors separately aligns with how these authors and texts are typically positioned within the discipline and enables more meaningful comparative analysis in this context. For DSRL, the demographic extraction process followed the same procedure; however, ethnicity was labelled using broad regional categories: North American, European, Australian, Asian, African and South American. These labels reflect general patterns of affiliation across a diverse academic field and support analysis of broader representational dynamics within the reading lists.

For all the datasets (i.e., AIMESRL, QSIERL and DSRL); to address cases where information was limited, supplementary details such as institutional affiliation and nationality were gathered from public databases and reputable websites. For example, determining the ethnicity and nationality of a political science scholar required consulting faculty pages, Wikipedia entries and publicly available interviews. Figure 3.1 illustrates the process of identifying ethnicity, demonstrating a case in which the author's nationality is explicitly indicated.

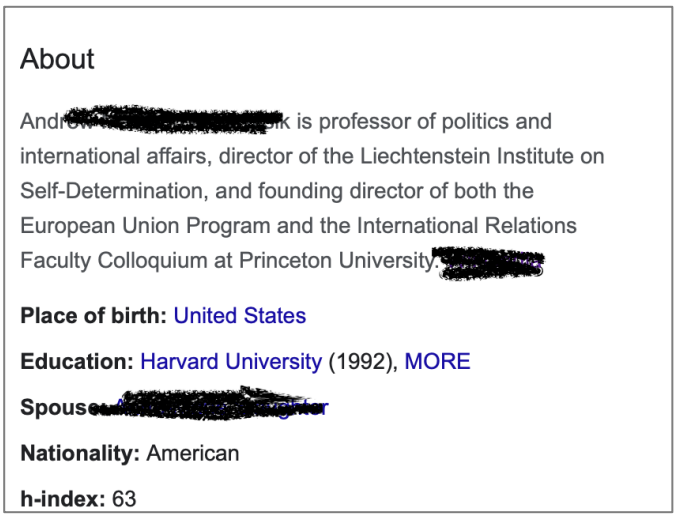

Figure 3.1: Example of public online information used for ethnicity labelling

Gender was occasionally inferred from contextual language for instance, a sentence such as "she is one of the leading authors in..." provided reasonable grounds to classify an author as female when no explicit gender identification was available. Figure 3.2 provides a clear example of gender identification, as the author is referenced using, she/her pronouns.

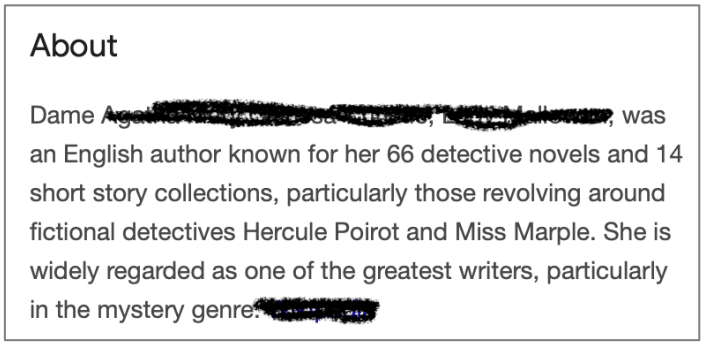

Figure 3.2: Example of public online information used for gender labelling

This triangulated approach, whilst not flawless, enabled the assignment of demographic categories with reasonable confidence and consistency across the datasets. Where no clear information about ethnicity could be identified, entries were similarly labelled as "unclear." While these categories are inevitably approximate, they were applied consistently to enable comparative analysis of representational patterns across the three reading list datasets.

3.3.3 Higher Education Online Course Materials: Core Learning Content Sources

Educational learning materials represent the most direct point of contact between institutional knowledge and student learning. Unlike news articles or reading lists, these materials directly shape how students understand concepts, interpret historical events and

form perspectives about society (Sleeter and Grant, 2008; Apple, 2014). This dataset focuses on the texts that students encounter during their learning process, where bias can have immediate and lasting impact on their educational experience.

3.3.3.1 The Primary Dataset: Higher Educational Materials

The primary educational materials dataset consists of higher educational materials, including lecture transcripts, learning notes and other study resources, all sourced from Open Educational Resources (OER) Commons[1] focus on learning materials in Humanities disciplines (OER-H). OER Commons is a globally accessible platform that provides a wealth of freely available educational content, allowing educators to modify and share resources under open licences.

3.3.3.2 Why OER Commons?

The original intention of this research was to analyse recorded lecture transcripts from the University of Leeds to examine pedagogical approaches and content delivery in higher education. However, access to these materials proved significantly limited due to institutional restrictions. Educational content housed on the University of Leeds's Minerva platform could not be utilised because permission and copyright restrictions prevented download for research purposes, as those materials are designated exclusively for internal teaching use. Consequently, even after seeking ethical review approval from the ethics committee and obtaining permission from the lecturers of these courses, only a small number of transcripts could be obtained, which was insufficient for a comprehensive analysis. This access barrier necessitated seeking alternative sources of educational content, leading to the adoption of OER Commons as the primary data source for this study.

This platform hosts a diverse range of educational materials, including textbooks, course syllabi, lecture videos, assessments, simulations and lesson plans, catering to educational needs from preschool to postgraduate levels. Key contributors to the OER Commons platform include over 70,000 resources from partners like Harvard University and the British Library, enabling tailored educational content creation. Additionally, MIT Open Courseware delivers comprehensive course materials for free, while Rice University's OpenStax provides peer-reviewed, openly licensed textbooks designed to reduce student costs. Other notable initiatives include the University of Minnesota's Open Textbook Library, which curates a broad selection of open-access textbooks and Carnegie Mellon University's Open Learning Initiative, which fosters resource sharing to enhance educational outcomes.

The resources extracted for this study from the OER-H data include materials provided by Yale Law School[2], the University of Michigan[3], the National Endowment for the

[1] https://oercommons.org/

[2] https://avalon.law.yale.edu

[3] https://quod.lib.umich.edu/lib/colllist/

Humanities[4], Lumen Learning[5], the Library of Congress[6], the United States National Archives[7] and Khan Academy[8].

3.3.3.3 Content Selection: Keyword-driven Extraction for Bias-relevant Materials

OER-H were sourced based on keyword searches available on the OER Commons platform. Keywords were selected to represent gender and ethnic bias in historical contexts, including terms like "gender roles", "racial discrimination", "colonial narratives" and "civil rights". Based on the search filters in OER Commons, "History" and "Arts and Humanities" were selected as the primary subjects, ensuring a focus on historical bias representation. The education level was set to "Graduate/Professional", aligning with university-level materials. The dataset size was determined based on content availability and the need for manageable manual validation. Twenty-five resources were extracted and each source was segmented into smaller text segments, resulting in a total of 290 distinct entries, each ranging from 150 to 250 words. This segmentation was intentionally designed to provide sufficient context for detecting subtle biases whilst remaining manageable for fine-tuning process.

All materials used in this study are licensed under Creative Commons Attribution-Non-Commercial-Share-Alike 4.0, allowing for their non-commercial use and adaptation, provided proper attribution is given.

The materials in OER-H are distinct because they are derived from academic contexts, specifically designed for university-level education. This makes them particularly suitable for investigating subtle and complex manifestations of bias in educational narratives, as these texts are often expected to be neutral or authoritative but may still perpetuate gender and ethnic stereotypes. Moreover, the dataset is novel in its inclusion of historical and autobiographical content that addresses critical societal issues, such as social movements, inequality and historical events.

3.3.3.4 Data Size

The OER-H dataset size reflects the intensive annotation methodology employed rather than data availability constraints. Each text segment underwent comprehensive multi-dimensional bias analysis, requiring detailed examination across representation, stereotype and linguistic categories. The annotation process demanded extensive human validation involving multiple annotators providing detailed justifications for each classification

[4] https://www.neh.gov

[5] https://lumenlearning.com

[6] https://www.loc.gov

[7] https://www.archives.gov

[8] https://www.khanacademy.org

decision, making the annotation process significantly more resource-intensive than simple binary classification tasks.

The 290 segments, derived from 25 educational resources and segmented into 150-250 word units, provided sufficient diversity across historical and humanities content whilst remaining manageable for the detailed annotation framework. This scale enabled comprehensive coverage of different educational contexts from lecture transcripts to instructional materials whilst ensuring that each instance received thorough analytical attention.

3.3.3.5 Domain Focus: Humanities Disciplines - Where Bias Often Hides in Plain Sight

Whilst discovering bias is not the main contribution of this work, the focus lies in developing robust tools that can help identify and analyse bias in educational contexts. The selection of Humanities disciplines materials was strategic: these domains offer more opportunities to encounter bias, which in turn enables the development of more effective detection tools and methodologies.

Humanities disciplines represent domains where bias is particularly prevalent yet often subtle (Foster, 1999; Loewen, 2007). Historical narratives frequently embed particular viewpoints whilst presenting themselves as objective accounts, making them ideal subjects for bias detection research (Wineburg, 2010). Humanities materials similarly contain implicit assumptions about culture, society and human behaviour that can reinforce existing stereotypes or marginalise certain perspectives (Said, 1978; Spivak, 2023).

Research in educational content analysis has consistently identified History and Humanities as disciplines where bias manifests through selective inclusion of voices, perspectives and interpretations (Christian-Smith and Apple, 1991; Romanowski, 1996). These fields often deal with contested narratives and multiple interpretations of events, making them particularly rich environments for training and validating bias detection systems. By working with content that contains diverse manifestations of the three bias types examined in this study, the resulting tools and methods can be developed to handle the full spectrum of educational bias detection challenges.

3.3.3.6 What This Data Reveals About Bias?

The OER-H dataset was designed to investigate the primary bias categories of gender and ethnicity bias, with representational, stereotypical and linguistic biases examined as secondary categories. The comprehensive approach recognises that learning materials often contain multiple, interconnected forms of bias that compound to influence student understanding (Sleeter, 2005; Gay, 2018).

1. **Representation: Whose Voices and Perspectives Shape Student Learning?**

Representation bias in educational materials manifests through the selection and prominence of different voices and perspectives within the content (Banks, 2004). This includes

examining whose historical accounts are privileged, which cultural perspectives are presented as normative and whose experiences are marginalised or omitted entirely. Research has shown that educational materials often prioritise dominant cultural narratives whilst minimising or excluding alternative viewpoints (Ladson-Billings, 2003; Au, 2009).

The investigation of representation bias in this dataset focuses on identifying patterns in whose knowledge is valued, whose experiences are centred in historical accounts and whose perspectives are treated as authoritative sources of information. This analysis extends beyond simple demographic counting to examine the quality and context of representation within educational narratives.

An example of representation bias in this data is the statement: "Despite the Fifteenth Amendment's failure to guarantee female suffrage, women did gain the right to vote in western territories, with the Wyoming Territory leading the way in 1869. One reason for this was a belief that giving women the right to vote would provide a moral compass to the otherwise lawless western frontier. Extending the right to vote in western territories also provided an incentive for White women to emigrate to the West…" [9] This demonstrates gender and racial representation bias by explicitly centring White women's experiences and motivations whilst marginalising the intersectional struggles of Black women and other women of colour in the suffrage movement.

2. Stereotype: Fixed Assumptions Embedded in Educational Narratives

Stereotype bias appears in educational materials through the reinforcement of oversimplified or generalised characteristics attributed to particular groups (Bigler and Liben, 2007). This can include assumptions about gender roles, cultural practices, or group capabilities that are presented as factual rather than socially constructed (Devine, 1989). Educational materials are particularly influential in perpetuating stereotypes because students often perceive them as authoritative and objective sources of information (Textbook League, 2000).

The analysis examines how groups are described, what characteristics are emphasised or assumed and whether materials challenge or reinforce existing social stereotypes. This investigation is crucial because educational content plays a significant role in shaping student attitudes and expectations about different social groups (Allport, 1954; Pettigrew and Tropp, 2006).

An example of stereotype bias in this data is the statement: "Following this defeat, many suffragists like Stanton increasingly replaced the ideal of universal suffrage with arguments about the virtue that White women would bring to the polls. These new arguments often hinged on racism and declared the necessity of White women voters to keep Black men

[9] Source link: https://courses.lumenlearning.com/wm-ushistory1/chapter/the-fifteenth-amendment/

in check". [10] This contains clear racial stereotypes and biological determinism, portraying White women as morally superior whilst depicting Black men as requiring control, thereby reinforcing both gender and racial hierarchies through stereotypical characterisations of virtue and threat.

3. Linguistic: Explicit and Implicit Bias Language in Educational Content

Linguistic bias involves the use of evaluative or emotionally charged language that influences reader perception whilst maintaining the appearance of objectivity (Fairclough, 2001). In educational materials, this often manifests through subtle word choices, metaphors and framing devices that guide student interpretation toward particular conclusions (van Dijk, 2015).

This type of bias is particularly insidious in educational contexts because students may not recognise the subjective nature of seemingly factual presentations (Luke, 1988). The analysis focuses on identifying language patterns that reveal implicit judgements, emotional framing, or persuasive techniques embedded within educational content. Below are examples of linguistic bias in OER-H:

Implicit example - This excerpt from lecture transcripts demonstrates subtle gender bias: "Each small group will then help its speaker to develop his or her argument. During the following class session give the principal speaker for each side an allotted amount of time to make his or her speech." [11] The phrase "his or her" reflects a binary understanding of gender, which may exclude non-binary individuals. While intending to include both male and female students, the language assumes all participants identify strictly as male or female, potentially marginalising non-binary students.

Explicit example - This historical text shows clear regional and political bias: "The controlling majority of the Federal Government, under various pretences and disguises, has so administered the same as to exclude the citizens of the Southern States, unless under odious and unconstitutional restrictions, from all the immense territory owned in common by all the States on the Pacific Ocean … By the disloyalty of the Northern States and their citizens and the imbecility of the Federal Government, infamous combinations of incendiaries and outlaws have been permitted …." [12]. The charged language ("odious", "infamous", "imbecility") portrays Northern States and the Federal Government in a highly negative light, reflecting the intense regional biases and political grievances typical of Southern secessionist rhetoric.

---

[10] Source link: https://courses.lumenlearning.com/wm-ushistory1/chapter/the-fifteenth-amendment/

[11] Source link: https://avalon.law.yale.edu/19th_century/csa_texsec.asp

[12] Source link: https://edsitement.neh.gov/lesson-plans/lesson-1-martin-luther-king-jr-and-nonviolent-resistance

3.3.3.7 Domain Transfer Validation Dataset

In addition to the primary OER-H dataset, a separate domain transfer validation dataset was constructed from text segments derived from recorded lecture transcripts from the University of Leeds master's degree programmes, specifically from the Artificial Intelligence MSc programme's Data Science and Programming for Data Science modules lecture transcripts (UoL-DSPDS-LT). The text segments were processed using the same segmentation approach as the OER-H, with each segment ranging from 150 to 200 words to maintain consistency across datasets, resulting in 223 text segments.

Proper ethical permissions were obtained from the module lecturers prior to accessing and utilising these materials for research purposes. The lecture materials were extracted from transcripts of 13 recorded lectures that were predominantly interactive in nature, featuring active engagement between lecturers and students rather than traditional one-way teaching delivery. This interactive format captures authentic classroom discourse, including spontaneous explanations, student questions and real-time pedagogical adjustments that may reveal implicit assumptions or biases not present in formal written materials (Cazden, 1988). Proper ethical permissions were acquired to use these lecture recordings for research purposes, with all materials subsequently anonymised to protect participant privacy.

UoL-DSPDS-LT serves a distinct methodological purpose: testing whether bias detection models trained exclusively on Humanities disciplines materials can effectively generalise to entirely different academic domains. The STEM lecture materials represent a critical test of cross-domain model performance. Technical disciplines are often assumed to maintain objectivity and neutrality, making them an ideal testing ground for examining whether bias detection methods developed on traditionally bias-prone subjects can identify more subtle manifestations of bias in seemingly neutral technical education (Harding, 1991; D'Ignazio and Klein, 2020).

Since OER-H have some general educational text such as classroom instructional materials, learning activities and pedagogical guidance, some examples are considered general educational text not domain specific include: "Arrange desks so that each team faces the other. Each team chooses three speakers, one to make the main points of the argument...", "Activity 1. Changes in the Franchise... Distribute the handout 'Examples of Changes in the Franchise'..." and "Essay: Ask each student to write a two-page essay that answers the lesson's guiding question..."

3.3.4 Complementary Perspectives on Bias in Higher Education

This study adopts a multi-dataset design to examine bias across different dimensions of online higher education. Rather than operating as sequential or interdependent stages, the three datasets of university news articles, reading lists and higher education learning materials provide complementary perspectives on how bias is embedded within institutional discourse, curricular structures and learning materials. Taken together, they offer a

comprehensive view of how knowledge is represented, structured and delivered within higher education contexts.

The analysis of university news articles focuses on institutional discourse, capturing how universities publicly construct narratives around academic achievement, authority and representation. This dataset provides insight into the symbolic and communicative practices through which institutions position scholars and disciplines. Patterns in representation and evaluative language reveal how particular identities and forms of knowledge are foregrounded or marginalised at the level of public-facing communication.

The reading list analysis shifts attention to the curriculum, examining how knowledge is formally organised within teaching contexts. By analysing authorship, demographic representation, and disciplinary coverage, this stage highlights which voices are prioritised within academic instruction. The inclusion of materials from both Western and Middle Eastern institutions introduces a comparative dimension, allowing for the identification of both shared and context-specific patterns of bias across different educational traditions.

The analysis of higher educational learning materials extends the investigation to the level of direct student engagement. This dataset enables the examination of how bias manifests within the content itself, including issues of representation, stereotyping and evaluative framing. It provides a detailed view of the materials through which knowledge is communicated and interpreted by learners.

Together, these three strands of analysis capture bias at distinct but complementary levels: institutional representation, curricular selection and instructional content. While each dataset offers a self-contained perspective, their combined analysis strengthens the overall explanatory scope of the study. By addressing multiple layers of the higher education ecosystem, the research moves beyond isolated observations to present a more comprehensive account of how bias operates across the production, organisation and dissemination of knowledge.

Below is a structured summary of the six datasets used in this integrated investigation, demonstrating how each provides unique analytical capabilities and methodological contributions to the comprehensive bias detection framework.

UoL-NA

- **Size:** 5,768 articles (2,174,466 words)
- **Data Contents:** News articles categorised by gender (see Chapter 4, section 4.4)
- **Institution:** University of Leeds
- **Primary Bias Categories:** Gender
- **Secondary Bias Categories:** Stereotyping, Representation and Linguistic
- **Dataset Usage:** Bias in text investigation, gender representation patterns and sentiment analysis in university news articles

AIMESRL

- **Size:** 212 reading items (4,142 words)
- **Data Contents:** Type, title, author name, author gender, author ethnicity, publisher and place of publication
- **Institution:** University of Leeds
- **Primary Bias Categories:** Gender and Ethnic
- **Secondary Bias Categories:** Representation
- **Dataset Usage:** Cross-cultural authorship diversity analysis and thematic content comparison

QSIERL

- **Size:** 73 reading items (936 words)
- **Data Contents:** Type, title, author name, gender, ethnicity and English availability
- **Institution:** Imam Mohammad Ibn Saud Islamic University
- **Primary Bias Categories:** Gender and Ethnic
- **Secondary Bias Categories:** Representation
- **Dataset Usage:** Cross-cultural authorship diversity analysis and thematic content comparison

DSRL

- **Size:** 679 reading items (15,274 words)
- **Data Contents:** Type, title, chapter title, author, chapter author, gender, ethnicity, publisher and place of publication
- **Institution:** University of Leeds
- **Primary Bias Categories:** Gender and Ethnic
- **Secondary Bias Categories:** Representation
- **Dataset Usage:** Interdisciplinary authorship diversity across disability studies curriculum

OER-H

- **Size:** 290 text segments (78,831 words)
- **Data Contents:** Text segment, type, bias class, bias category, bias keywords/phrases and justification
- **Institution:** Multiple via OER Commons
- **Primary Bias Categories:** Gender and Ethnic
- **Secondary Bias Categories:** Stereotyping, Representation and Linguistic
- **Dataset Usage:** Training and building multi-dimensional bias detection model development and validation

UoL-DSPDS-LT

- **Size:** 223 text segments
- **Data Contents:** Text segments only
- **Institution:** University of Leeds
- **Primary Bias Categories:** Gender and Ethnic
- **Secondary Bias Categories:** Stereotyping, Representation and Linguistic

- **Dataset Usage:** Cross-domain model generalisation testing from humanities to STEM contexts

3.4 The Annotation Process: From AI Detection to Human Evaluation

Bias detection in educational texts presents unique challenges that demand human expertise and cultural awareness. Research has consistently shown that annotation can introduce bias through mismatches between annotator populations and the data being analysed (Hovy and Prabhumoye, 2021; Sap et al., 2021), making careful annotator selection critical for educational bias detection. Human perspectives are essential for mitigating bias-related issues in AI systems, though they also introduce challenges associated with cognitive biases inherent in human decision-making (Gautam and Srinath, 2024).

However, human annotation is not without significant limitations. Human biases inevitably shape annotation outcomes, particularly because annotators' backgrounds such as cultural norms, beliefs and disciplinary training can influence their interpretation of textual cues (Shah, D.J. et al., 2018). Research demonstrates that annotators with different attitudes and identities systematically bias their judgements, especially in sensitive domains like bias detection (Sap et al., 2021). The challenge of human label variation due to annotator perspective biases creates substantial inconsistencies, with studies showing that what one individual perceives as biased may differ significantly based on their cultural context and personal experiences (Plank, 2022).

These inconsistencies have led researchers to explore AI-assisted annotation approaches. LLMs have demonstrated remarkable capabilities in annotation tasks, often providing more consistent and scalable solutions than human annotators alone (Gilardi et al., 2023). Studies comparing human and AI annotators have found that LLMs can achieve comparable or superior performance to human annotators on various text classification tasks whilst maintaining greater consistency across annotations (Gilardi et al., 2023; Törnberg, 2023; Divya Venkatesh et al., 2024; Pangakis and Wolken, 2024b; Rouzegar and Makrehchi, 2024; Wang, X. et al., 2024; Calderon et al., 2025; Nemkova et al., 2025).

Gilardi et al. (2023) demonstrated that ChatGPT's zero-shot accuracy exceeds that of crowd workers by approximately 25 percentage points on average across multiple annotation tasks including relevance, stance and topic detection. Similarly, Divya Venkatesh et al. (2024) found that comparing human text classification performance with LLMs using eye-tracking revealed that simple ML models can perform at par or better than untrained LLMs and non-expert humans for domain-specific text classification tasks. Nemkova et al. (2025) reported that state-of-the-art LLMs achieve superior performance in human rights violation detection tasks in multilingual social media data, particularly excelling in precision and recall metrics. Rouzegar and Makrehchi (2024) showed that LLMs can enhance text classification through active learning frameworks that integrate human annotators and LLMs, demonstrating improved accuracy and efficiency. Wang, X. et al. (2024) found that human-LLM collaborative annotation through effective verification of LLM

labels significantly enhances text classification tasks whilst maintaining cost-effectiveness. Pangakis and Wolken (2024b) demonstrated that knowledge distillation approaches using LLM-generated training labels can achieve performance comparable to human-annotated datasets in automated text classification.

AI annotation offers several advantages: it eliminates individual cognitive biases, provides consistent decision-making criteria and can process large volumes of text efficiently (Ziems et al., 2024).

However, the optimal approach appears to be neither purely human nor purely AI-based annotation. Research increasingly supports hybrid methodologies that combine the strengths of both approaches (Huang, F. et al., 2023). The most effective strategy involves initial AI annotation to establish consistent baseline classifications, followed by human validation to capture cultural nuances and contextual understanding that AI might miss (Kuzman et al., 2023b). This hybrid approach leverages AI's consistency and scalability whilst incorporating human expertise for quality assurance and cultural sensitivity.

Building on this annotation approach, two different annotation frameworks were employed to annotate the datasets used in this study: the university news articles and the OER Commons learning materials. The following sections describe each framework in detail, outlining the specific methodologies, quality control measures and validation approaches applied to ensure reliable and consistent bias detection across both educational contexts.

The following section presents three annotation approaches: the first uses human annotators first before comparing their judgements against LLM-generated annotations (see section 3.5.2), the second relies solely on human annotators (see section 3.5.3) and the third employs LLMs to generate initial annotations which human annotators then review and correct (see section 3.6).

3.5 UoL-NA Annotation Framework

3.5.1 Overview of Annotation Approaches

This annotation process encompassed two primary tasks: Linguistic bias specifically subjectivity bias classification and sentiment classification. The subjectivity bias classification task focused on determining whether an article contained biased language. This process involved binary labelling, where articles were marked as either biased or non-biased. In contrast, sentiment classification assigned each article a sentiment category positive, negative, or neutral based on the overall tone and framing of the content. While bias classification aims to identify subjective or slanted reporting, sentiment classification focuses on evaluating the emotional tone conveyed within the text. An article classified as biased may exhibit preferential language, omission of alternative viewpoints, or a lack of neutrality, whereas sentiment classification captures whether the content presents individuals, events, or topics in a favourable, critical, or neutral light. The distinction between

these tasks is crucial, as an article may carry a strong sentiment without being explicitly biased and vice versa.

A combination of LLM-based methods and human annotations was employed to ensure a comprehensive evaluation of bias in university news articles. For the analysis of linguistic bias, subjectivity bias classification was performed using both LLM-based approaches and human annotation, whereas sentiment classification relied solely on human annotation. The following section provides a detailed description of each process and its implementation.

3.5.2 Annotation Process for Linguistic (Subjectivity) Bias

3.5.2.1 Human Annotations for Subjectivity Bias in UoL-NA

Before starting the human annotation process, a key priority was ensuring that annotators had a clear understanding of the context of university news articles. Since bias in this domain tends to be more subtle compared to mainstream media, annotators were provided with detailed instructions to help them accurately identify patterns of bias. Unlike traditional news outlets, where bias is often overt such as through political framing or emotionally charged language, bias in academic news is more likely to appear in who is represented, how they are portrayed and the thematic focus of coverage.

As an example of mainstream media bias from the MBIC dataset (Spinde et al., 2021), which serves as a primary data source for media bias reporting:

> "People keep getting sick and the economy is cratering, but for Republicans, hating women is still a major priority: republican governors in Texas, Ohio and even Maryland are trying to use the coronavirus crisis as an excuse to block women from getting abortions, claiming that ending a pregnancy is a medical procedure that must be rescheduled."

In contrast, bias in university news articles tends to be more subtle and implicit, often manifesting through gender representation and thematic framing rather than overtly charged language. For instance, a biased university news article might present achievements differently based on gender:

> "Professor XX, a pioneering female scientist, has balanced her groundbreaking research with raising a family, demonstrating that women can succeed in both academia and personal life."

While this statement appears positive, it reinforces traditional gender roles by linking the professor's academic success to her family responsibilities, a framing that is rarely applied to male academics. This representational bias subtly perpetuates stereotypical

expectations and highlights the different ways men and women are portrayed in academic news compared to mainstream media bias, which is often more explicit and politically charged.

This study focuses specifically on gender and ethnic biases, with particular attention to subjectivity and representation bias expressed through language. Representation bias emerges when certain groups are overrepresented, underrepresented, or stereotypically portrayed in coverage, potentially reinforcing unequal visibility or harmful narratives. Subjectivity bias, on the other hand, occurs when reporting includes personal opinions, evaluative language, or emotionally charged framing that may influence readers' perceptions rather than presenting information in a neutral, fact-based manner. In this study, annotators were instructed to closely examine how language and tone shape these portrayals, for example, by noting patterns where male academics were predominantly associated with career achievements, while female academics were more frequently framed in relation to community engagement or family-oriented themes.

For the annotation process, Amazon Mechanical Turk (MTurk) [13] was utilised through Amazon SageMaker's Ground Truth feature. A random subset of 110 records was assigned to annotators, with each news article being reviewed by three independent annotators to ensure high reliability. To maintain consistency and accuracy, annotators were provided with detailed guidelines on identifying subjectivity bias in the text and instructed to classify each article as either biased or non-biased. Annotators were compensated at a rate of $0.480 per task and were primarily based in the EU (London) region. Out of the random sample 88 where classified biased and 22 classified as not biased.

3.5.2.2 Inter-Annotator Agreement and Reliability for Subjectivity Bias in UoL-NA

To assess the reliability of human annotations, inter-annotator agreement was measured for both bias classifications. Since bias in university news articles is often subtle and context-dependent rather than explicitly stated, achieving a high level of agreement among annotators was particularly challenging. Two key metrics were used to evaluate agreement: percentage agreement and Cohen's Kappa scores.
For bias classification, percentage agreement measured the proportion of cases where all annotators assigned the same label. However, percentage agreement alone does not account for agreement occurring by chance, which is why Cohen's Kappa (κ) scores were also computed. Cohen's Kappa is a more robust measure of reliability that adjusts for random agreement, calculated using the following equation:

$$k = \frac{P_o - P_e}{1 - P_e}$$

[13] https://www.mturk.com

where:

- P_o (Observed Agreement) is the proportion of cases where annotators assigned the same label.
- P_e (Expected Agreement by Chance) is the probability of agreement occurring randomly.

The Kappa values range from -1 to 1, where values close to 1 indicate strong agreement, values near 0 suggest agreement is no better than chance and negative values imply systematic disagreement. Given the contextual and interpretative nature of bias detection, Cohen's Kappa provides a more accurate measure of annotation reliability than simple percentage agreement.

The overall percentage agreement was 52.3%, indicating that in slightly more than half of the cases, annotators reached a consensus. However, Cohen's Kappa scores revealed inconsistencies, with Annotator 1 & Annotator 2 achieving only slight agreement (0.165), Annotator 1 & Annotator 3 showing almost no agreement (0.012) and Annotator 2 & Annotator 3 having a negative agreement score (-0.122), suggesting active disagreement. The negative Kappa score highlights significant differences in how annotators perceived bias, reinforcing the difficulty of identifying subjectivity bias in academic media, where language tends to be institutionally neutral yet subtly unbalanced in representation.

3.5.2.3 LLMs Zero-Shot Annotations for Subjectivity Bias in UoL-NA

This study employed two LLMs OpenAI's GPT-3.5-turbo[14] and Google's Bard (now rebranded as Gemini) [15] for zero-shot automated annotation of UoL-NA. These models were used to assess the capability of LLMs to perform accurate data annotation; therefore, their outputs were systematically evaluated against a subset of 110 human-annotated data articles to measure agreement and reliability. These models were evaluated based on their accuracy, scalability and cost-effectiveness in performing subjectivity bias classification.

GPT-3.5-turbo was chosen due to its extensive use in NLP applications and its capacity to efficiently process large-scale text annotation tasks (Gilardi et al., 2023; Huang, F. et al., 2023; Kuzman et al., 2023a). The model's performance was assessed in terms of annotation accuracy, particularly in detecting biased language within the dataset. Additionally, its ability to maintain consistent classifications across different articles was examined to evaluate its reliability.

Similarly, Google Bard was utilised as a comparative model to identify differences in annotation performance. The evaluation focused on its reliability in bias detection and how

[14] https://docs.aimlapi.com/api-references/text-models-llm/openai/gpt-3.5-turbo

[15] https://gemini.google.com

its classification outputs compared to those of GPT-3.5-turbo. This comparison aimed to highlight the strengths and limitations of each model in the specific context of bias analysis within academic media.

GPT-3.5-turbo was used to annotate the subset from UoL-NA, with a cost of $0.002 per 1,000 tokens via OpenAI's Application Programming Interfaces (API). In contrast, Google Bard provided annotations for the same sample at no cost, as it was accessed directly through the chat interface.

Both LLMs were given the same prompt to ensure consistency in their annotations:

> Prompt: "You are a helpful AI that can identify all types of subjectivity bias in text. You will read a text from the news that is provided to you and reply with either 'biased' or 'unbiased'."

3.5.2.4 Performance Evaluation of LLMs Zero-Shot Annotations for Subjectivity Bias in UoL-NA

GPT-3.5-turbo and Google Bard were evaluated to determine how well they performed in bias detection when compared to MTurk human annotations. The models' performance was measured using accuracy, precision, recall and F1-score, as shown in Table 3.1. The following is a brief explanation of each metric, accompanied by its mathematical formula:

- **Accuracy**

Accuracy measures the overall proportion of correctly classified instances out of all predictions. It is suitable for balanced datasets but can be misleading when one class dominates.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Where:

- *TP* = True Positives
- *TN* = True Negatives
- *FP* = False Positives
- *FN* = False Negatives

- **Precision**

Precision evaluates the proportion of correctly predicted positive observations among all predicted positives. It is particularly important when the cost of false positives is high.

$$Precision = \frac{TP}{TP + FP}$$

- **Recall**

Also referred to as sensitivity or the true positive rate, recall measures the proportion of actual positive cases that were correctly identified by the model.

$$Recall = \frac{TP}{TP + FN}$$

- **F1-Score**

The F1-score is the harmonic mean of precision and recall. It is a balanced metric that is especially useful when a trade-off between precision and recall is necessary.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Model	Accu-racy	Preci-sion	Recall	F1-score
GPT-3.5-turbo	30.2%	95.2%	20.6%	33.9%
Google Bard	40.4%	94.4%	35.1%	51.1%

Table 3.1: Performance of zero-shot GPT-3.5-turbo and Google Bard against MTurk annotations for subjectivity bias in UoL-NA data

Based on Table 3.1 the results indicate that Google Bard outperformed GPT-3.5-turbo in terms of accuracy, recall, and F1-score. This suggests that Google Bard was slightly better at identifying bias in university news articles. However, GPT-3.5-turbo exhibited a higher precision, meaning that when it identified bias, it was more likely to be correct than Google Bard.

Despite these differences, both models showed low recall values, meaning they frequently failed to detect biased articles. This limitation suggests that while LLMs can be useful in confirming bias when detected, they are not yet reliable for comprehensive bias annotation tasks, particularly in domain-specific contexts like higher education news.

3.5.3 Annotation Process for Sentiment Analysis in UoL-NA

For sentiment classification, MTurk was used to manually annotate 46 entries in UoL-NA. Annotators were instructed to classify each article as positive, negative, or neutral, with a specific focus on the sentiment directed towards the individual mentioned in the article. The compensation rate and annotator region remained the same as in the previous bias annotation task. To ensure reliability, each article was annotated by three independent annotators, and inter-annotator agreement was calculated to assess the consistency of sentiment classifications.

To enhance accuracy and consistency, annotators were provided with detailed instructions along with examples illustrating different sentiment classifications. A positive sentiment typically highlighted achievements or recognitions (e.g., "Dr. XX was awarded the prestigious research grant"), while a negative sentiment often involved controversies or criticism (e.g., "Dr. XX's controversial statement sparked criticism"). Neutral sentiment was assigned to factual statements without emotional tone or evaluative language (e.g., "Professor XX presented his findings."). This guidance helped ensure that sentiment classification was aligned with the study's focus on gender and ethnic biases analysis, particularly in understanding how individuals are framed within university news articles.

Of the total randomly selected sample, the annotators classified 4 items as neutral, 19 as positive, and 23 as negative.

3.5.3.1 Inter-Annotator Agreement and Reliability for Sentiment Analysis in UoL-NA

Two key metrics were used to evaluate agreement: percentage agreement and Cohen's Kappa scores. For sentiment classification, the level of agreement among annotators was even lower, indicating that identifying sentiment in UoL-NA was highly subjective. The percentage agreement was 38.9%, meaning that in the majority of cases, annotators assigned different sentiment labels to the same article. This was further reflected in the Cohen's Kappa scores, which remained low across all annotator pairs. Annotator 1 & Annotator 2 showed almost no agreement (0.020), Annotator 1 & Annotator 3 achieved a fair level of agreement (0.263), and Annotator 2 & Annotator 3 had only slight agreement (0.080).

These results suggest that sentiment annotation was more challenging than bias classification, likely due to differences in how annotators perceived emotional tone and framing. Unlike mainstream news, where sentiment is often explicit, university news articles tend to use neutral and professional language, making it difficult to determine whether the tone is purely factual, subtly positive, or subtly negative. Additionally, sentiment in university news may not always be directed at an individual but rather at the institution, a research breakthrough, or a policy change, further complicating classification.

3.6 OER-H Bias Annotation Framework

Learning from the challenges encountered in the UoL-NA annotation process, a comprehensive annotation framework was developed specifically for OER-H that incorporated several methodological improvements designed to enhance annotation quality and consistency.

The framework addressed the platform limitations observed in the UoL-NA by utilising Prolific [16] for OER-H annotation. Research demonstrates that advanced platforms with ex-ante vetting controls significantly outperform general crowdsourcing platforms on quality measures (Palan and Schitter, 2018). Unlike the UoL-NA which relied on MTurk's general workforce, the OER-H framework required annotators with specific educational qualifications.

The enhanced recruitment strategy implemented stringent selection criteria such as annotators had backgrounds aligned with the domain and cultural context of the educational content. This approach follows established research indicating that annotation quality significantly improves when annotator expertise matches task requirements (Poesio and Artstein, 2005).
Another critical improvement was the adoption of a three bias classification labels ('Bias', 'Not Bias', 'Potential Bias') rather than the binary approach used for UoL-NA. This methodological advancement addressed the inherent ambiguity in educational bias detection, where content may exhibit subtle or context-dependent bias that defies binary classification.

The 'Potential Bias' category serves a crucial function in educational content analysis, capturing instances where bias exists but requires additional context for definitive classification (Aroyo and Welty, 2015). Research in annotation theory emphasises that acknowledgment of ambiguity through graduated classification schemes improves annotation reliability and reduces artificially inflated agreement scores (Plank, 2022). This approach recognises that educational bias often manifests through nuanced representational patterns rather than explicit discriminatory language.
Moreover, this framework implemented a systematic, multi-stage guideline development process designed to address the annotation inconsistencies observed in the news articles study. Following established annotation best practices (Pustejovsky and Stubbs, 2012), the development process comprised iterative cycles of guideline creation, pilot annotation, disagreement analysis, and guideline refinement.

The iterative approach began with initial guideline drafts based on bias detection literature, followed by pilot annotation sessions with small subsets of educational materials. Annotator disagreements were systematically analysed to identify unclear instructions or

[16] https://www.prolific.com

ambiguous category boundaries. Research indicates that iterative refinement significantly improves both annotation quality and inter-annotator agreement (Fort et al., 2011). Guidelines were progressively refined through multiple rounds until achieving stable annotation patterns and adequate agreement thresholds.

One of the key innovations was the mandatory requirement for annotators to provide detailed justifications for all classification decisions. This approach addresses research findings that requiring explanations significantly improves annotation quality by forcing annotators to engage more carefully with the content (Pavlick and Kwiatkowski, 2019). The justification requirement served multiple purposes: enhancing annotator attention and engagement, providing transparency in decision-making processes, enabling quality assessment beyond simple label agreement, and creating rich explanatory data for model training. Research demonstrates that explanation requirements reduce careless annotation and improve overall data quality (Hsueh et al., 2009).

The framework leveraged Prolific's advanced quality control features, including the ability to return incomplete or inadequate annotations to workers for revision. This capability addresses a significant limitation observed in the news articles study, where annotations could not be efficiently corrected. The revision mechanism operated through systematic quality checks, clear feedback on annotation deficiencies, opportunity for annotators to revise and resubmit work, and rejection protocols for annotations not meeting quality standards. This approach aligns with research showing that feedback loops and revision opportunities significantly improve annotation quality in complex tasks (He et al., 2023).

Lastly, a key addition to this framework was the incorporation of AI-assisted annotation capabilities, creating a hybrid human-AI annotation system that leveraged the strengths of both approaches. The AI annotation component utilised LLMs to provide preliminary bias assessments and highlight potentially problematic content segments for human review, with the AI system serving as a first-pass filter that identified clear cases of bias or non-bias while flagging ambiguous instances for detailed human evaluation. This integration addressed several limitations observed in purely human-based annotation processes while maintaining the nuanced judgment capabilities essential for educational bias detection, and research demonstrates that AI-human collaboration in annotation tasks can significantly improve both efficiency and consistency while reducing annotator fatigue (Park, J.K. et al., 2024).

3.6.1 Initial Annotation Using GPT-4o

The initial annotation of OER-H with 290 entries was performed using OpenAI's GPT-4o[17], a state-of-the-art LLM. The aim was to automate the initial bias detection process to streamline human validation. Each text segment underwent an initial classification for bias, where the model assigned one of three possible bias decisions: 'Bias', 'Not Bias', or

[17] https://openai.com/index/hello-gpt-4o/

'Potential Bias'. The classification as 'Bias' indicated that the text clearly contained elements of stereotyping, misrepresentation, or biased language. A 'Not Bias' decision signified that the text was deemed neutral, with no discernible bias. In cases where bias was less explicit but might be inferred depending on the context or reader's perspective, the text was marked as 'Potential Bias', suggesting that a deeper analysis would be necessary to confirm the presence of bias.

For text segments marked as 'Bias' or 'Potential Bias', GPT-4o further categorised the bias into three specific types: 'Stereotype', 'Representation' and 'Linguistic'. Stereotype bias referred to the generalisation of roles or attributes based on gender, ethnicity, or social class. Representation bias identified instances where certain groups were underrepresented or portrayed in a limited, stereotypical manner. Linguistic bias focused on the use of language that either favoured one group over others or conveyed an inherently biased tone. One or more of these categories were applied to each 'Bias' or 'Potential Bias' segment. Additionally, specific keywords or phrases contributing to the decision were highlighted, and a brief justification was provided to explain the reasoning behind the annotation, ensuring transparency in the model's decision-making process, regardless of the bias decisions. Using GPT-4o for the initial annotation allowed for the efficient processing of the dataset while providing a structured foundation for human validation in the subsequent stages. The annotation was conducted using the following prompt:

> "Prompt: You are a bias-detection expert reviewing higher educational materials. Your task is to analyse the given text and determine:
>
> Bias Decision: Classify the text as one of the following:
>
> Bias: The text clearly contains elements of stereotyping, misrepresentation, or biased language. Not Bias: The text is neutral, with no discernible bias. Potential Bias: The bias is less explicit but might be inferred depending on context or reader perception, requiring deeper analysis. If biased, categorise it under one or more of the following categories: Stereotype: Generalising roles, attributes, or behaviours to a specific gender, ethnic group, or social class. Representation: Under-representation or limited portrayals of certain genders, ethnicities, or social groups, leading to exclusion or marginalisation. Linguistic: Use of language that favours or promotes one group over others, or a tone that reflects implicit bias. Highlight specific biased words or phrases found in the text. Provide a clear explanation supporting your classification".

3.6.2 Human Annotation using Prolific

As previously mentioned, the full OER-H data of 290 instances was initially annotated using GPT-4.0 to generate preliminary bias classifications. Of these, 140 instances were subsequently re-annotated and validated by human annotators. To ensure the accuracy and reliability of the bias annotations, a human validation phase was conducted using Prolific, a platform known for its high-quality participant selection. The selection criteria for annotators were carefully designed to ensure domain expertise and familiarity with the societal and cultural contexts relevant to the texts under review. All annotators were native English speakers and held degrees in Arts and Humanities, Education, Social Sciences, or History, providing them with the necessary academic background to critically assess bias. Additionally, annotators were based in the UK or the US to align with the cultural context of the educational materials being reviewed.

The annotation task was divided into phases to ensure that the instructions were clear and that annotators followed them correctly. In the first phase, three annotators were selected: a 53-year-old male with a Social Sciences degree from the UK of White ethnicity, a 26-year-old male with a Social Sciences degree from the UK of Asian ethnicity, and a 55-year-old male with an Arts and Humanities degree from the UK of White ethnicity. After this initial round, the selection criteria were revised to incorporate a more diverse group of annotators, broadening the range of perspectives. Subsequent annotation rounds placed greater emphasis on diversity in gender and ethnicity while maintaining the requirement for native English proficiency due to the language of the data. Examples of annotators in later phases included a 30-year-old male with a Social Sciences degree from the UK of Asian (Pakistani) ethnicity, a 20-year-old female with a Social Sciences degree from the USA of African (Nigerian) ethnicity, and a 46-year-old female with an Arts and Humanities degree from the UK of African (Ghanaian) ethnicity. This iterative approach ensured a balance between domain expertise, cultural context, and diversity in perspectives, allowing for a more comprehensive validation of the bias annotations.

To maintain consistency and reliability, annotators followed detailed annotation instructions, which outlined the bias classification process, justification requirements, and examples of well-formed and poorly justified annotations. The annotation guidelines and examples were provided to all annotators. Annotators' consent was obtained, and they were provided with full instructions, including details on the use of the data they annotated. Before participation, annotators were presented with the following statement: "By participating in this study and completing you agree that your responses and any provided data may be used as part of a research study. The data may be included in academic publications, presentations, or reports. You have the option to decline by exiting the study at any time. Do you consent to participate and allow your responses to be used for research purposes?"

To assess the consistency of human annotations, inter-annotator agreement was measured using both percentage agreement and Cohen's Kappa Scores. The percentage agreement, which reflects the proportion of cases where all three annotators provided the same

classification, was 45.71%, indicating that less than half of the annotated instances had full agreement. This suggests a level of subjectivity in bias classification, where different annotators may have interpreted the criteria differently. To further quantify agreement beyond chance, Cohen's Kappa Scores were calculated for each pair of annotators. The agreement between Annotator 1 and Annotator 2 resulted in a Kappa score of 0.35, suggesting fair agreement, while the score between Annotator 1 and Annotator 3 was higher at 0.52, indicating moderate agreement. The agreement between Annotator 2 and Annotator 3 was 0.39, also falling within the fair range. These results highlight the challenge of maintaining consistency in bias classification, particularly in complex domains such as history content, where interpretations of bias can be nuanced and context dependent.

Following the annotation process, an analysis of the inter-annotator agreement labels revealed an imbalance in the distribution of bias classifications. 'Not Bias' was the most frequently assigned label with 59 instances, followed by 'Potential Bias' with 45 instances, while 'Bias' was the least assigned category with only 36 instances. This imbalance suggested a tendency among annotators to exercise caution in definitively labelling a text as biased, often opting for a potential bias classification when uncertainty was present. To ensure a more balanced dataset for model training and evaluation, we addressed this issue by selecting an equal number of instances from each class. Specifically, we sampled 36 instances of each class, creating a balanced training and evaluation set of 108 human-annotated instances. This approach ensured that the model was trained on an equal representation of each bias category, preventing any skewed learning patterns that could arise from class imbalances and improving the model's ability to generalise across different bias classifications. Moreover, a confidence score was assigned based on annotator agreement.

The confidence scores were calculated based on the level of agreement amongst annotators. For each text segment, we measured the proportion of annotators who agreed with the initial decision (Bias, Not Bias, or Potential Bias). The majority of segments were reviewed by three annotators, although in some cases a fourth annotator was included to provide an additional perspective.

Confidence was defined as follows:

- High confidence (100%; 4/4 or 3/3 agreement): all annotators agreed with the bias decision.
- Moderate confidence (75% or 66.67%; 3/4 or 2/3 agreement): one annotator disagreed with the majority.
- Low confidence (50%; 2/4 agreement): annotators were evenly split.

3.6.3 Reliability and Independent Human Evaluation of LLM Annotations

Given that the annotators were reviewing and re-annotating LLM-generated annotations, a specific protocol was used to ensure that they engaged critically with each text segment rather than merely confirming automated outputs. This section sets out detailed evidence

that annotators applied independent judgement and thorough analytical reasoning throughout the task.

First, annotators were provided with detailed task instructions that explicitly prohibited the use of AI tools and underscored the requirement for manual, critical analysis. The instructions stated clearly that "this task must be completed manually without using any form of AI tools" and further specified that annotators who felt unable to complete the task accurately should exit rather than submit uncertain or incomplete responses. This functioned as a filtering mechanism, helping to ensure that only annotators confident in their ability to provide independent and well-reasoned judgements took part in the study. Crucially, the instructions also ruled out vague responses such as "I agree with the annotation" and required that justifications "reference specific parts of the text, focusing on the reasoning behind your decision". This meant that annotators were not able merely to endorse the initial annotations, but instead had to articulate a clear, evidence-based rationale for each decision.

Second, the annotator justifications provide strong evidence of genuine critical engagement rather than routine agreement with LLM outputs. Analysis of the 440 justification comments associated with the 140 annotated text segments, produced mainly by three annotators, with a fourth annotator involved in 20 cases, shows substantial variation in annotator reasoning. A recount indicates that 83 segments (59.3%) contain at least one explicit marker of challenge, contrast, qualification or disagreement in the annotators' justifications. These cases do not consist only of direct rejection of the original label. In many instances, annotators broadly accepted the initial bias decision but qualified it by using terms such as however, but or although, offering alternative reasoning, disputing the proposed bias category, or arguing that the justification overstated the evidence. In other cases, annotators explicitly disagreed with the decision or challenged the assumptions underlying the interpretation of the text. Across all 440 comments, explicit agreement markers appear in 82 comments (18.6%), whereas explicit disagreement markers appear in 30 comments (6.8%).

Taken together, these patterns show that annotators were not merely endorsing prior classifications. Rather, they were engaging critically with both the label and the reasoning behind it, often refining, narrowing or reframing the interpretation even where they did not wholly reject it. The examples below further illustrate the depth and independence of annotator analysis.

Example 1: Disagreement on bias classification and historical context

For a text segment containing the term "negroes" in a historical context, annotators provided markedly different assessments:

Annotator 1: "The identification of "negroes" is correct; however, this is 'representation' and 'linguistic' bias. It is not correct that this is 'stereotyping' because there is no generalising taking place in the text."

Annotator 2: "I do not agree that this is bias as historical language should not be corrected to suit today's language."

This example shows that annotators did not simply accept the initial classification. Annotator 1 accepted that the passage raised a concern but challenged the specific bias category and supplied an alternative categorisation grounded in textual analysis. Annotator 2 went further by questioning the broader premise that historical language should be evaluated according to contemporary linguistic norms. The disagreement therefore operated at two levels: the classification itself and the methodological framework underpinning the judgement.

Example 2: Disagreement on contextual omission

A text segment focusing on voting rights for white males elicited divergent responses:

Annotator 1: "I would change the decision to 'potential bias' as it is extremely likely that other sections of the scheme of work would focus on disenfranchised groups."

Annotator 2: "The text is very clear in direction and focuses on white males; there is no need to expand this to other people who did not have the right to vote, as this is not what the text asks for. Excluding a certain group of people from a task does not create bias."

Annotator 1 did not dismiss the concern entirely but argued that the classification should be moderated in light of the wider educational context. Annotator 2, by contrast, rejected the premise that selective historical focus necessarily constitutes bias. This illustrates a recurring pattern in the dataset: annotators often agreed that a passage required scrutiny, but differed over the severity, basis or contextual interpretation of the bias judgement.

Example 3: LLM's justification critique

Annotators frequently challenged not only the bias classifications but also the quality and logic of the justifications themselves. The following examples, drawn from different segments, illustrate this pattern:

Annotator: "The justification reads as the opinion of the writer rather than fact. The information is factual and therefore cannot be bias."

Annotator: "I feel the justification is written as opinion rather than fact. The original text states exactly what happened."

These comments show that annotators were assessing the evidential quality of the reasoning, not merely the decision outcome. In particular, they distinguished between factual description and evaluative interpretation and questioned whether some justifications relied too heavily on subjective opinion. Such comments provide strong evidence of active analytical engagement with the annotation process itself.

Example 4: Nuanced Disagreement on Language in a Slavery-Era Text

For a historical text containing references to slavery, one annotator provided particularly nuanced analysis:

Annotator: "The phrasing of the justification given is obscure. There is a slight suggestion of bias in the fact the narrator suggests they bought a 'negro' solely as a favour, but still seem to regard the person as an asset by whose decision to leave they could incur a loss. Without other clearer material about the narrator's attitude to slavery, this bias seems only potential."

This justification demonstrates several layers of critical reasoning. First, the annotator critiques the original justification as insufficiently clear. Secondly, they identify a subtle bias cue not fully captured in the original explanation. Thirdly, they qualify the judgement by arguing that the evidence supports only potential bias rather than a definitive classification. Finally, they recognise that the interpretation depends on contextual information not available in the segment itself. This level of careful qualification is inconsistent with passive verification and instead reflects thoughtful manual judgement.

Example 5: Disagreement on historical description, representation and factual framing

For a segment describing "peace between the whites and the Sioux", annotators again differed in whether the wording should be treated as neutral description or as a potentially biased framing:

Annotator 1: "A fine decision, but in general, educational resources should avoid using such terminology as 'the whites' or 'the blacks' in favour of more nuanced details."

Annotator 2: "There is limiting portrayals of genders, ethnicities, or social groups in the text segment because it does not provide the full context of the power imbalance. This means there is some potential bias because it is not clear that the white people had much more power and force to meet their demands. Due to the limiting portrayals, there is a representation issue in the text segment."

Annotator 3: "I feel the justification is written as opinion rather than fact. The original text states exactly what happened."

Annotator 4: "The phrase 'peace between the whites and the Sioux' oversimplifies the complex and often coercive nature of treaty negotiations."

This case further illustrates the subjective and interpretative character of the task. Annotator 1 accepted the general decision but refined the reasoning towards pedagogical language choice. Annotator 2 argued that omission of the power imbalance created a representation problem. Annotator 3 challenged the justification as overly opinion-based, whilst Annotator 4 argued that the wording itself obscured coercion by presenting the

event as neutral peace-making. The disagreement therefore concerned not only whether bias was present, but also whether bias arose through terminology, omission or interpretative overreach.

Overall, the evidence presented in this section demonstrates that the annotation process involved substantive manual labour, independent judgement and critical analytical engagement. Annotators did not simply verify LLM outputs; rather, they provided detailed, reasoned justifications that frequently qualified, challenged, refined or reinterpreted the original annotations and their supporting rationales. The diversity of perspectives, the sophistication of the reasoning and the recurring patterns of disagreement all indicate that the annotations reflect genuine human judgement about the complex and subjective phenomenon of bias in educational texts. At the same time, this analysis should be understood as evidence that annotators carried out the task in the manner requested, using their own judgement and expertise. It does not imply that bias annotation is free from subjectivity, nor that the resulting labels constitute unquestionable ground truth. Rather, the annotations should be understood as informed human judgements within an inherently interpretative task. As Example five illustrates, this subjectivity can arise even when annotators engage carefully with the same passage, as different readers may reasonably weigh historical context, wording and implied meaning in different ways.

3.6.4 GPT-4o Annotations Versus Human Annotations

A comparative analysis between the annotations generated by GPT-4o and those validated by human annotators revealed several important insights. Overall, GPT-4o demonstrated an 81.43% agreement with human annotators, suggesting strong alignment in many cases. The model's classifications were more evenly distributed across 'Bias', 'Potential Bias' and 'Not Bias', whereas human annotators leaned more towards classifying text as 'Not Bias', followed by 'Potential Bias' and 'Bias'. This difference suggests that GPT-4o was more likely to flag bias compared to human annotators, who exercised more caution and tended to classify texts as neutral. For bias categories GPT-4o annotated 47 instances of Stereotype, while the human annotation found 44. For Representation, GPT-4o identified 97 cases compared to the human's 83, and for Linguistic GPT-4o marked 80 instances, whereas the human annotation recorded 78. Overall, GPT-4o consistently showed slightly higher counts across all bias types.

One possible explanation for this discrepancy is that GPT-4o may be overly sensitive to bias indicators, flagging text as biased even in cases where human annotators did not perceive it as such. Human annotators, being more aware of the broader context and intent behind the text, were often more discerning in their classifications. They were able to account for nuances that the model struggled with, particularly in cases where bias was subtle or dependent on historical, social, or linguistic context. In contrast, GPT-4o's heightened sensitivity may have led to an over-identification of bias, potentially classifying neutral or contextually justified statements as biased.

The confidence score, which was assigned based on annotator agreement, provided a quantitative measure of annotation reliability, helping to distinguish between cases where bias classification was straightforward and cases where it was more ambiguous. Segments with higher confidence scores (e.g., 90%) indicated strong agreement among human annotators, suggesting that these instances were less subjective and easier to classify. GPT-4o tended to align more closely with these high-confidence annotations, reinforcing its effectiveness in identifying explicit biases. However, for text segments with lower confidence scores (e.g., 50% or below), where annotators disagreed on the classification, GPT-4o also showed poor justification for the classification. These lower-confidence cases often involved subtle biases, implicit associations, or complex historical and cultural references, where contextual interpretation played a crucial role in classification.

Additionally, the requirement for human annotators to provide justifications for their classifications, whether they agreed with the model or not, proved instrumental in refining the dataset and improving transparency. These justifications, combined with confidence scores, helped uncover cases where bias detection was subjective or where additional context was necessary to make an informed decision. The findings highlight the strengths and limitations of GPT-4o in bias detection, demonstrating that while it aligns well with human judgment in many cases, its tendency toward over-sensitivity underscores the importance of human expertise in interpreting and contextualising bias. The confidence score played a key role in evaluating annotation reliability, ensuring that the dataset remained transparent and well-balanced for model training and evaluation.

3.7 Methodological Limitations and Ethical Considerations

Several methodological limitations and ethical considerations warrant acknowledgement:

3.7.1 Reading Lists Annotation

The manual demographic annotation of reading lists introduces potential classification biases, particularly in the treatment of ethnicity. Categorising authors into broad ethnic groups risks oversimplifying complex cultural identities and can obscure intersectional factors such as mixed heritage, migration history, or self-identification. Moreover, such classifications often rely on external assumptions based on names or biographical information, which may not accurately reflect an author's lived experience or chosen identity. This can lead to misrepresentation, reinforcing rigid or reductive categories and potentially introducing systemic bias into downstream analyses.

3.7.2 UoL-NA Annotation Framework Ethical Considerations

3.7.2.1 MTurk Annotation

One of the main methodological limitations of this annotation framework was the inter-annotator agreement reliability. On one hand, the low inter-annotator agreement in bias classification suggests that annotators may have interpreted the guidelines differently or applied subjective judgments inconsistently. Some bias indicators, such as who is featured in an article or how their achievements are framed, require contextual understanding that varies between individuals. This discrepancy highlights the need for more refined annotation guidelines, enhanced training sessions, and potentially a consensus-based labelling strategy to improve the consistency of human annotations.

On the other hand, the reliability of annotations obtained through MTurk has been a subject of debate, particularly for tasks requiring nuanced understanding, such as bias and sentiment analysis in academic texts. Several studies have highlighted concerns regarding the quality and consistency of MTurk annotations in such contexts. One significant issue is the variability in annotator performance. Research indicates that while MTurk can provide quick and cost-effective data, the quality of annotations can be inconsistent, especially for complex tasks that demand a deep understanding of context and subtlety. For instance, studies have found that MTurk workers may not consistently produce high-quality data for tasks involving subjective judgment or specialised knowledge (Keith et al., 2017).

Another concern is the potential for deceptive responses due to financial motivations. The compensation structure on MTurk might incentivise workers to complete tasks rapidly rather than accurately, leading to superficial or inattentive annotations. This tendency is particularly problematic in domains like academic media, where detecting subtle biases requires careful and thoughtful analysis (Fort et al., 2011). Furthermore, the assumption that MTurk's 'Master' workers, those with higher approval ratings, would provide superior data quality has been challenged. Empirical evidence suggests that requiring Master qualifications does not necessarily result in more reliable annotations for tasks assessing personality and cognitive abilities. This finding implies that even experienced MTurk workers may struggle with tasks that require nuanced understanding, such as identifying subtle biases in text (Rouse, 2019).

In the context of this study, which focuses on detecting subtle gender and ethnic biases in university news articles, the limitations of MTurk annotations become particularly pronounced. The low inter-annotator agreement observed suggests that MTurk workers may lack the necessary contextual knowledge and sensitivity to consistently identify nuanced biases in academic content. This unreliability raises concerns about the dependability of MTurk for annotating complex, context-dependent tasks in specialised domains.

3.7.2.2 Zero-shot LLMs Annotation

LLM-based annotations face several challenges that impact their reliability and consistency. One key issue is the variability in predictions across different prompts. In some instances, GPT-3.5-turbo has even produced two different responses simultaneously, prompting the user to choose between them. This approach, which serves as a feedback test for a random subset of users, highlights the inconsistency in the model's outputs and raises concerns about its reliability in sensitive applications like bias detection.

Another significant challenge is the limited domain-specific understanding of these models, particularly in academic contexts. University news articles often contain specialised terminology and nuanced expressions that may not be well-represented in the general training data of these LLMs. As a result, the models can struggle to accurately identify subtle forms of bias or correctly interpret the tone of the content, further compromising the quality of automated annotations.

3.7.3 OER-H Annotation Framework Ethical Considerations

3.7.3.1 Consistency and Fairness in LLMs

The use of LLMs for annotating bias in educational content raises several ethical concerns, particularly around the consistency and fairness of these models' classifications. While LLMs are invaluable for automating large-scale tasks, their tendency to misclassify or inconsistently label sensitive content presents significant ethical challenges especially when dealing with nuanced topics like bias.

One of the most significant ethical issues observed in this study was the inconsistency in GPT-4o's classification of bias. In some instances, the model revised its own classification when asked to re-evaluate the same text segment, indicating a lack of reliability. For example, GPT-4o initially classified the phrase "I purchased my wife Meg" as 'Not Bias', justifying it as a familial description. However, when prompted to reevaluate, GPT-4o corrected its decision to 'Bias', recognising the dehumanising nature of the language, which reduces an individual to property and reflects harmful stereotypes tied to slavery. This shift in classification illustrates the model's uncertainty and inconsistency in recognising complex biases, especially those embedded in historical or societal contexts.

Similarly, in another instance, a segment discussing the "expulsion of African Americans" from a community was initially labelled as 'Not Bias', as GPT-4o failed to identify the racial discrimination implied in the phrase. Upon re-analysis, the model correctly reclassified the text as 'Bias', acknowledging the offensive nature of the language and its historical significance. These examples raise concerns about the model's reliability in consistently detecting bias, as its decisions fluctuate depending on the context or upon repeated evaluation. Such inconsistencies are ethically problematic, especially when the model is used

in sensitive applications like educational content annotation, where fairness and precision are paramount.

Another challenge encountered was the issue of double classification, where GPT-4o would assign multiple conflicting labels to a single text segment. In one case, the model simultaneously labelled a segment as both 'Bias' and 'Potential Bias', leaving the user to decide which classification to adopt. This not only dilutes the precision of the annotation but also blurs the line between distinct types of bias. The model's difficulty in distinguishing between intertwined biases can lead to over-generalisation, misrepresenting the actual nature of the bias in question. This raises further ethical concerns about the accuracy and transparency of automated bias detection, as such ambiguity could lead to misinterpretation or, worse, reinforce existing biases.

Despite these inconsistencies, GPT-4o demonstrated an overall accuracy of 81.43%, aligning well with human judgment in cases of explicit bias, such as overt stereotyping or discriminatory language. However, more nuanced cases such as implicit biases in representation often presented challenges for the model. These fluctuating classifications highlight the ethical importance of human oversight in the annotation process. In this study, human annotators from Prolific were involved in reviewing and validating the model's annotations to ensure consistency and fairness. Human review serves as a critical safeguard, helping to address the model's limitations and ensuring that subtle forms of bias are not overlooked or misclassified. While LLMs offer substantial benefits in automating large-scale tasks like bias annotation, their occasional inconsistencies and double classifications underscore the need for a multi-step review process. Ensuring that annotations are reviewed, validated, and reconsidered when inconsistencies arise is essential to maintaining ethical integrity. LLMs are powerful tools, but they are not infallible, and without human oversight, they risk reinforcing biases or failing to detect subtle discriminatory patterns.

3.7.3.2 Human Annotator Perspectives

While human annotators play a critical role in ensuring accurate data annotations, they are not immune to personal bias. Individual interpretations of language, social issues, and evolving societal norms can often reflect personal beliefs and experiences, which may influence the annotation process. This is particularly important in sensitive topics like bias detection, where subjective judgment can lead to inconsistent outcomes, even though we chose a more reliable annotation platform and selected annotators based on their educational backgrounds, which are closely related to the themes in our data. For instance, in our study, a segment was flagged by GPT-4o as 'Bias' due to its use of gendered language:

> "Each small group will then help its speaker to develop his or her argument. During the following class session, give the principal speaker for each side an allotted amount of time to make his or her speech".

The phrase "his or her" was identified as reflecting a binary understanding of gender, potentially excluding non-binary individuals. When the segment was reviewed by human annotators, most agreed that this language was exclusionary and recommended a more inclusive alternative. However, one annotator expressed a differing opinion, which reflected personal beliefs rooted in a traditional, binary view of gender:

Annotator 1: "The phrase 'his or her' reflects a binary understanding of gender, which may exclude non-binary individuals." (Agreed with GPT-4o's decision).

Annotator 2: "The use of 'his or her' favours or promotes non-transgender persons. It is biased under 'linguistic'. 'Their' would be a more inclusive alternative." (Agreed with GPT-4o's decision).

Annotator 3: "Gender is binary. Someone who identifies as non-binary is still either male or female." (Disagreed with GPT-4o's decision and annotated this as 'Not Bias').

While two annotators recognised the exclusionary nature of the phrase, Annotator 3's response reflected a personal bias rooted in traditional, binary views of gender. This highlights the challenge of human subjectivity, even when working with annotators selected for their expertise. Despite the shared academic background, differences in personal beliefs influenced their interpretations, underscoring the ethical complexities involved in human-led annotations.

Another issue arose in another segment flagged for potential bias by GPT-4o. The following text was part of an educational resource discussing Malcolm X's views and the Nation of Islam's beliefs:

> "Why does Malcolm X call white people 'devils'? Why does the Nation of Islam support a racially segregated American society? What solution does Malcolm X propose and how would this provide what blacks most need? How does Malcolm X defend the Nation of Islam from the criticism that they preach black supremacy and violence?"

GPT-4o flagged this segment as 'Potential Bias', noting that the phrasing of the questions could introduce bias or oversimplify complex ideas, potentially shaping how students perceive Malcolm X and the Nation of Islam. The question "Why does Malcolm X call white people 'devils'?" was highlighted for focusing on one of the most controversial aspects of Malcolm X's rhetoric, which may reinforce negative stereotypes without offering the necessary context. Similarly, the question "Why does the Nation of Islam support a racially segregated American society?" could be perceived as presenting the Nation of Islam's beliefs in a negative light without exploring the historical reasons behind their stance.

The annotators' responses varied significantly:

Annotator 1: "The text focuses on selected soundbites and not the overall text and deeper detail issued by Malcolm X." (Agreed with GPT-4o's decision as 'Potential Bias').

Annotator 2: "I feel the justification is trying too hard to protect a certain image of Malcolm X. The fact remains that he said certain things that are perceived a certain way, and I don't feel it necessary to explain that to today's audience." (Disagreed with GPT-4o's decision and annotated this as 'Not Bias').

Annotator 3: "These are leading questions and would lead to misrepresentation and bias on the views of Malcolm X. The questions suggest that Malcolm X is a bad character by posing these questions without prior and additional context." (Disagreed with GPT-4o's decision and annotated this as 'Bias').

Once again, the responses illustrate the subjective nature of bias detection, even among annotators with expertise in the subject matter. While some identified bias in the leading nature of the questions, others believed that the broader educational context provided sufficient mitigation against potential misrepresentation. This highlights the inherent complexities and challenges of relying on human annotators for bias evaluation, as individual perspectives, interpretive frameworks, and personal assumptions can significantly shape their judgments.

To enhance the accuracy and reliability of our assessment while minimising the impact of individual bias, we introduced a fourth annotator for additional evaluation. Annotator 4's response was as follows:

Annotator 4: "It's clear that the text is part of a larger educational resource. It will be expanded upon by further content that will deal with the terminology and historical context." (Disagreed with GPT-4o's decision and annotated as 'Not Bias').

By incorporating an additional annotator, the study aimed to minimise the influence of individual subjectivity and achieve a more balanced and reliable consensus. This approach ensured that the final annotations reflected a broader and more nuanced understanding of bias, particularly in cases where initial disagreements emerged. The inclusion of a fourth annotator ultimately helped settle the classification, providing further validation of the decision. However, this method was applied selectively, only in instances where annotations varied significantly across different evaluators.

While human annotators play a crucial role in refining and validating automated bias detection systems, managing subjectivity requires careful selection of diverse perspectives and structured oversight. These strategies are essential for maintaining fairness and consistency in the annotation process, particularly in complex and interpretive areas such as bias identification. Despite these efforts, complete agreement was not always possible, as reflected by only moderate inter-annotator agreement scores. This highlights the broader challenge of operationalising bias detection, a process that presents difficulties not only for automated models but also for human evaluators. Bias annotation is fundamentally shaped by

individual and cultural perspectives, and absolute objectivity is rarely attainable. Whilst this study sought to mitigate these issues through diverse annotator pools and structured guidelines, the limitations of subjectivity should be considered when interpreting both human and model performance.

3.7.3.3 Confidence Scores Based on Annotator Agreement

Confidence scores were used to provide a consistent summary of the level of agreement amongst annotators for each text segment. This approach offered a practical and transparent way to represent variation in annotator agreement across the dataset, making it possible to distinguish between cases that attracted strong consensus and those that remained more contested.

At the same time, this method has important limitations. Confidence scores reduce nuanced interpretative disagreement to a numerical measure of agreement and therefore cannot capture the full reasoning behind annotators' decisions. In the context of bias annotation, disagreement does not necessarily indicate error or poor-quality judgement; in many cases, it reflects the inherently subjective, contextual, and contestable nature of the task. Confidence scores should therefore not be interpreted as a direct measure of objective correctness.

This limitation is particularly relevant when the dataset is used for model training. A model trained on these scores may learn to over-weight majority judgements as though they represent definitive truth, rather than recognising that some cases remain open to reasonable disagreement. For this reason, the confidence score should be understood as a pragmatic indicator of consensus rather than a claim to final or unquestionable ground truth. Although imperfect, this approach provides a usable way to retain information about agreement levels whilst acknowledging the interpretative complexity of bias annotation.

3.8 The Methodological Framework of the Upcoming Chapters

This research employs a deliberately progressive approach to methodological complexity, beginning with foundational techniques and advancing towards sophisticated AI systems. This strategy ensures robust validation at each stage whilst building towards comprehensive bias detection capabilities suitable for real-world educational applications. The subsequent chapters will extend this progression, continuing the investigation using the datasets introduced in this chapter as follows:

Chapter 4: Foundation building with statistical and traditional NLP approaches (e.g., sentiment analysis) applied to UoL-NA. This foundational stage establishes baseline patterns of institutional bias whilst validating fundamental detection approaches. Traditional methods including frequency analysis, lexicon-based classification, and sentiment analysis provide interpretable results that form the empirical foundation for subsequent methodological advances.

Chapter 5: The methodology expands to incorporate unsupervised learning, cross-cultural validation on university module reading lists (i.e., AIMESRL and QSIERL) and intersectional bias analysis (i.e., DSRL). Topic modelling and demographic analysis techniques enable investigation of gender and ethnic representation patterns across different cultural and institutional contexts. This stage validates methodological transferability whilst introducing complexity through cross-cultural comparison and thematic analysis.

Chapters 6-7: Focusing on OER-H and UoL-DSPDS-LT, the research culminates in advanced AI methodologies including fine-tuned LLM, RAG (Lewis, P. et al., 2020) and hybrid systems. These sophisticated approaches address the limitations of earlier methods whilst maintaining interpretability through hybrid architectures.

Chapter 8: Design, prototyping and testing of a web application as a practical validation for OER-H. The final methodological stage transitions from research validation to practical tool through web applications. This implementation phase demonstrates the real-world applicability of developed methods whilst providing validation of their effectiveness in educational contexts.

3.9 Conclusion

This chapter established the foundation for investigating bias in higher education online content. It introduced the conceptual framework for identifying various bias categories, particularly those related to gender and ethnicity, and presented the selected data sources alongside the frameworks used for bias annotation across these datasets. By combining three distinct bias types with six complementary datasets spanning university news, reading lists, and direct learning materials the research develops a comprehensive framework for understanding how bias operates within the online higher educational ecosystem.

Furthermore, this chapter outlined the methodological framework that guides the subsequent chapters. The next chapter builds this foundation further by employing statistical and NLP techniques, such as sentiment analysis, applied to the UoL-NA dataset.

Chapter 4

4 Investigating Bias in University News Articles

4.1 Introduction

This chapter builds on the findings previously presented in (Bin Shiha et al., 2023; Bin Shiha et al., 2024), which were published as part of an earlier stage of this research. While the structure and methodology remain largely consistent with the original papers, some analyses have been revised, expanded or re-run due to an update in the dataset size. The analysis in this chapter is based on a larger or more complete version of the dataset, which has resulted in slight differences in the reported results compared to the original publication. Moreover, this chapter builds on the previous one, which introduced the dataset, by applying a range of NLP methods to investigate bias within the UoL-NA data.

4.1.1 Context and Importance of University News Articles

In this chapter, we investigate gender bias within the domain of university news articles, highlighting three primary forms of impact that such bias can have: 'Linguistics' 'Stereotype' and 'Representation' biases. Representational impact refers to the unequal representation and visibility given to individuals based on gender, which may lead to disparities in opportunities for recognition and engagement within the academic community (Wagner et al., 2015). Stereotype emerges through the stereotypical portrayal of genders, potentially reinforcing outdated norms and overlooking the diversity and complexity of gender identities. This bias not only perpetuates gender stereotypes but also systematically disadvantages female and non-binary individuals by limiting their representation in positive, authoritative, or diverse roles within academic news content (Moss-Racusin et al., 2012). Linguistic bias refers to the subtle ways in which reporting through word choice, phrasing, and overall stylistic tone can systematically favour one gender over another in university news coverage. This includes disproportionate use of emotionally loaded adjectives, differential framing of competence or authority, emphasis on personal roles (e.g., 'mother' vs. 'scientist'), or shifts in formality depending on gender. Such bias can shape how readers perceive individuals undermining their professional credibility, reinforcing stereotypes, and shaping institutional culture. This work is motivated by the belief that addressing these biases can lead to more inclusive and equitable representations in university news, thereby challenging and shifting the deep-rooted stereotypes that persist in society.

University news articles serve a dual role: they function as internal communications for staff and students while also influencing external perceptions of the institution. Covering topics such as faculty achievements, research breakthroughs, student initiatives, and policy

changes, these articles contribute to shaping the university's identity. Universities often present themselves as neutral and authoritative sources of knowledge, creating the perception that their news coverage is objective and unbiased (Hartley, M. and Morphew, 2008). However, this neutrality is not absolute; academic media, like all forms of institutional discourse, is shaped by power structures that determine which voices and perspectives receive prominence. As Rathbun and Turner (2012) argue, academic development practices and resources reflect dynamics of power, control, and authority, challenging the assumption that academic texts and by extension, university news can be truly neutral.

Bias in university-affiliated media can take multiple forms, including gender representation bias (e.g., disparities in coverage of male and female academics) and linguistic bias (e.g., gendered framing of achievements). Unlike mainstream media, which operates under commercial pressures and editorial independence, university news is typically managed by internal communications teams that prioritise institutional branding and stakeholder interests (Zerfass et al., 2017). This can result in overrepresentation of certain groups (e.g., senior male academics) while marginalising others (e.g., women and early-career researchers). Additionally, university news may exhibit sentiment bias, where articles are overwhelmingly positive to reinforce the institution's prestige (Bin Shiha et al., 2023; Bin Shiha et al., 2024).

Understanding these biases is crucial, as universities are influential institutions that shape public discourse, research priorities, and educational policies. Investigating bias in university news articles not only contributes to NLP research in bias detection but also promotes fairer and more inclusive institutional narratives.

4.1.2 Objectives of the Chapter

- **Establish baseline automated bias detection:** Classify subjectivity and biased language in university news articles using zero-shot classification with DistilBERT (Raza et al., 2022), pretrained on the Media Bias Annotation Dataset Including Annotator Characteristics (MBIC) (Spinde et al., 2021). Evaluate the pretrained model's performance against human annotations to establish baseline accuracy without requiring domain-specific training.
- **Compare zero-shot classification approaches:** Systematically evaluate the reliability, consistency, and scalability of a pretrained language model (DistilBERT) against large language models (OpenAI's GPT-3.5-turbo and Google's Bard, now rebranded as Gemini) for detecting gender-biased language patterns in university news articles.
- **Quantify gender representation disparities:** Apply distributional frequency analysis using lexicon-based NLP approaches to identify distinct linguistic vocabulary that differentiates male- and female-oriented coverage. Use chi-square tests (McHugh, 2013) to establish empirical baselines for systematic bias assessment and categorise articles into gendered groups.
- **Identify stereotypical language patterns:** Detect and categorise family and career language patterns in gendered articles using lexicon-based approaches, examining differential emphasis on professional achievements versus personal

attributes, family responsibilities versus career accomplishments, and leadership qualities versus collaborative characteristics.

- **Analyse sentiment and framing differences:** Implement the Facebook BART-Large-MNLI model[18] to produce quantitative sentiment scores across gendered articles. Compare zero-shot sentiment detection with human annotations to assess how male and female subjects are framed within university news coverage.

4.1.3 Contributions of The Chapter

This chapter contributes to NLP bias detection research by presenting the first investigation of gender bias in university news articles. This addresses the main contributions outlined in Chapter 1 Section 1.3.4.3.

4.2 Datasets

This study utilises UoL-NA dataset. The dataset underwent systematic preprocessing to prepare it for the analysis. Further details regarding data collection methodology and annotation process are provided in Chapter 3, Section 3.3.1.

4.2.1 External Datasets

In addition to the main dataset, several external resources were utilised to support the analysis and evaluation of bias within university news. These external datasets are listed below.

4.2.1.1 MBIC – A Media Bias Annotation Corpus

The MBIC (Spinde et al., 2021) is a widely used resource for bias detection in news articles. It comprises a diverse collection of news content sourced from multiple media platforms, including HuffPost, MSNBC, USA Today, and other prominent outlets.
For this study, an augmented version of the MBIC corpus was used, which was extended by Raza et al. (2022) to encompass a broader spectrum of biases. This expanded dataset includes additional annotations capturing biases related to gender, race, ethnicity, education, religion, and language, as identified in existing literature (Gaucher et al., 2011; Matfield, 2016). The extension enhances the corpus's diversity and comprehensiveness by incorporating manual annotations, ensuring a more nuanced representation of different bias types.

The corpus consists of 3,700 entries, each classified as either biased or non-biased, based on initial crowdsourced annotations. It includes various metadata features such as news snippets, URLs, source publications, topics, demographic attributes (age, gender, and education), biased words, and corresponding bias labels. This corpus serves as a crucial

[18] https://huggingface.co/facebook/bart-large-mnli

benchmark for evaluating bias detection models and analysing patterns of bias in news reporting.

4.2.1.2 Lexical Datasets for Gender and Thematic Analysis

In addition to the primary dataset, this study incorporates lexical datasets from Dacon and Liu (2021), which provide structured lists of gendered and thematic words. These datasets facilitate a systematic analysis of gender representation in university news articles by examining the presence and frequency of specific terms related to gender identity, career roles, and family associations.

- **Male Possessive Words:** A dataset containing 230 words associated with male possessive nouns, including terms such as gentlemen, chairman, policeman, landlord, and others that are traditionally linked to male-associated roles and identities.
- **Female Possessive Words:** A dataset comprising 235 words related to female possessive nouns, including landlady, businesswoman, mother, goddesses, and similar terms reflecting female-associated roles.
- **Career Words:** A dataset featuring 162 words related to career and professional roles, encompassing both gender-specific (e.g., actress, businessman, waitress) and gender-neutral (e.g., lawyer, engineer, nurse) occupational terms.
- **Family Words:** A dataset consisting of 195 words related to familial roles and relationships, including mother, father, family, orphan, and other terms frequently associated with family and caregiving contexts.

4.3 Zero-Shot Bias Detection in UoL-NA using PLM

This study employed PLM DistilBERT fine-tuned on the MBIC dataset to detect potential bias patterns in UoL-NA. DistilBERT is a distilled version of BERT, built using a teacher-student framework where the large BERT model acts as the 'teacher', and the smaller student model learns to mimic it by compressing its layers while retaining most of its linguistic understanding. This architecture allows DistilBERT to capture contextual word relationships crucial for identifying subtle linguistic patterns while being significantly faster and more efficient than the original BERT.

DistilBERT was selected for this task due to its computational efficiency and strong performance in text classification tasks. Compared to BERT, DistilBERT retains 97% of its accuracy while being 60% faster and requiring 40% fewer parameters (Sanh et al., 2019), making it a practical choice for large-scale bias detection in UoL-NA. The MBIC dataset used for fine-tuning encompasses a diverse range of biased and non-biased news snippets, ensuring that the model learns the linguistic structures associated with bias in news reporting (Sanh et al., 2019). Although bias in this dataset often represents political discourse rather than educational news, this model was chosen because it was trained on a large-scale news dataset. Importantly, while the majority of news sources come from

mainstream media, some articles are related to education, making it a relevant choice for this study.
While the DistilBERT model was originally designed to identify general media bias rather than gender-specific bias, research demonstrates that models trained for one type of bias detection can effectively transfer their learned representations to detect other forms of bias through transfer learning (Saunders and Byrne, 2020; Zhao, J. et al., 2020; Lu et al., 2024). Studies have shown that bias detection approaches can exhibit cross-domain transferability, where detecting one bias form (e.g., general media bias) may contribute to identifying others (e.g., gender bias) through shared linguistic patterns and framing mechanisms (Lu et al., 2024). Transfer learning approaches have been successfully applied to bias detection tasks, demonstrating that models can adapt their bias identification capabilities across different domains and contexts (Saunders and Byrne, 2020). The model's training on media bias detection provides a foundation for identifying systematic differences in how subjects are presented and framed in news content, which can be adapted to recognise gender-specific patterns in university communications.

When a news article is input, the text is first tokenised into word pieces (e.g., 'professor', 'inspires' and 'leading expert'). DistilBERT then creates contextual embeddings, meaning each word's representation depends on the words around it. This contextual understanding is crucial because gender bias in language is often contextual rather than found in isolated words (Hitti et al., 2019).

Consider these examples from the study context:

- "Professor Smith, a leading expert in quantum physics, presented the findings…"
- "Professor Johnson, a mother of two, shared her inspiring journey…"

Here, both articles describe academics, but the framing differs: the male professor is described by professional expertise, whilst the female professor is characterised by personal and familial roles. DistilBERT's classification head processes these contextual embeddings and predicts whether the article reflects biased or non-biased framing patterns. Each article receives a label (biased or not biased) and a probability score from 0 to 1 (e.g., 0.87 for high confidence or 0.52 for uncertain classification). In the example above, the article emphasising family roles for a female academic would likely receive a biased classification with higher confidence, whereas the expertise-focused article would be classified as not-biased. Experiments were conducted in a Jupyter Notebook on a MacBook Pro (Apple M2 Pro, 32 GB RAM).

4.3.1 Performance Evaluation of DistilBERT for Zero-Shot Bias Classification

To assess bias detection in university news articles, DistilBERT's classifications were evaluated against human annotations collected through MTurk. A randomly selected subset of 110 articles was annotated independently by the model and by three MTurk workers, providing a ground truth for measuring model reliability (see Chapter 3, Section 3.5 for the detailed annotation framework).

To evaluate the performance of the classification model, six commonly used metrics were employed: accuracy, precision, recall, F1-score, logarithmic loss (log loss) and area under the receiver operating characteristic curve (AUC-ROC). Each metric captures different aspects of model effectiveness and contributes to a more nuanced understanding of predictive performance, particularly in classification tasks where class distributions may be imbalanced.

The following section provides a brief explanation of log loss and AUC-ROC, along with their mathematical formulations. For the definitions and formulas of accuracy, precision, recall and F1-score see Chapter 3, Section 3.5.2.4.

- **Log Loss**

Log loss quantifies the uncertainty of predictions based on their probabilistic outputs. It penalises incorrect predictions more when they are made with high confidence. Lower log loss indicates better model performance.

$$Log\ Loss = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)]$$

Where:

- $y_i$is the true binary label for instance i
- p_i is the predicted probability for the positive class
- N is the total number of instances

- **AUC-ROC**

The ROC curve plots the true positive rate against the false positive rate at various thresholds, illustrating the trade-off between sensitivity and specificity. The AUC quantifies overall performance; values closer to 1.0 indicate better discrimination between classes. Figure 4.1 presents a generic example of an ROC curve, where the curve lies above the diagonal baseline, indicating better-than-random classification. The shaded area represents the AUC, which summarises the model's overall discriminative performance.

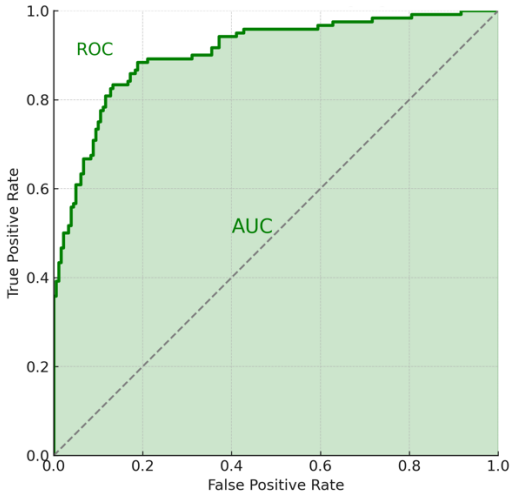

Figure 4.1: Example of ROC-AUC

Based on Table 4.1 the DistilBERT model demonstrated moderate accuracy, achieving 68.8%, meaning that it correctly classified most instances. The model's precision was high, indicating that when it flagged an article as biased, it was highly likely to be correct. However, its recall suggests that it did not capture all instances of bias, possibly missing some subtle or implicit cases. The F1-score reflects a strong balance between precision and recall, making DistilBERT a promising tool for bias detection in university news articles.

Despite its strengths, the log loss value of 0.457 suggests that the model had only moderate confidence in its predictions. More notably, the AUC-ROC score of 0.539 indicates that its ability to distinguish between biased and non-biased articles was only slightly better than random chance. This result underscores the challenge of bias detection in academic reporting, where language may be more subtle and nuanced than in general news sources.

A comparison of DistilBERT's predictions with MTurk human annotations further confirmed that while the model performed reasonably well, it struggled with domain-specific linguistic patterns, highlighting the need for further fine-tuning on university news datasets.

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**	**Log Loss**	**AUC-ROC**
DistilBERT	68.8%	89.9%	73.2%	80.7%	0.457	0.539

Table 4.1: DistilBERT performance metrics of bias detection on UoL-NA data

4.3.2 Comparative Analysis of UoL-NA Bias Detection Using LLMs and PLM

To fully assess the effectiveness of these models, a comparison was made between DistilBERT, GPT-3.5-turbo and Google Bard. The results, shown in Table 4.1, include the performance of DistilBERT. The LLMs (i.e., GPT-3.5-turbo and Bard) are shown in Table 4.2 below, and their performance here is compared against MTurk human annotations as the benchmark. These results highlight significant differences in accuracy, precision, recall, and F1-score.

Model	Accuracy	Precision	Recall	F1-score
GPT-3.5-turbo	30.2%	95.2%	20.6%	33.9%
Google Bard	40.4%	94.4%	35.1%	51.1%

Table 4.2: Performance of zero-shot GPT-3.5-turbo and Google Bard against MTurk annotations for subjectivity bias in UoL-NA data

The findings demonstrate that DistilBERT outperformed both GPT-3.5-turbo and Google Bard, particularly in recall, F1-score, and overall accuracy. While GPT-3.5-turbo and Google Bard had high precision, their low recall values indicate that they missed a significant number of biased articles, making them unreliable for bias detection. DistilBERT's recall of 73.2% suggests that it was more effective at identifying biased instances, though it still had limitations in distinguishing between subtle forms of bias.

The AUC-ROC score of 0.539 for DistilBERT highlights that while it performed better than random chance, it still requires further domain-specific training to improve its ability to differentiate between biased and non-biased articles.
Overall, these results indicate that DistilBERT is currently the best-performing model for bias detection in university news but still has room for improvement. The LLMs, despite their precision, are not reliable for comprehensive bias detection due to their low recall and accuracy scores. These findings suggest that bias detection in higher education news requires fine-tuned models trained on domain-specific data, rather than relying on general-purpose language models.

4.4 Gender Representation in UoL-NA through Distributional Frequency Analysis

This section outlines the methodology employed to analyse gender representation in UoL-NA, focusing on the classification of articles based on gendered terms and pronouns. The primary objective was to quantify and assess the distribution of gender-related language in academic media, particularly investigating whether male-oriented, female-oriented, or neutral representations were dominant in university news content.

To achieve this, the study adopted a lexicon-based approach, systematically identifying gendered language and categorising articles accordingly. The subsequent statistical evaluation provided insights into potential gender imbalances present in the dataset.

4.4.1 Lexicon-Based Approach for Gender Frequency Analysis

A lexicon-based approach was employed to systematically identify gendered terms within UoL-NA. This methodology is widely used in bias detection research due to its ability to efficiently classify textual data based on predefined lexicons.

The gender classification lexicon, derived from the datasets in Dacon and Liu (2021) work, included:

- **Male-associated terms:** 'he', 'his', 'him', 'himself', 'Mr.', 'Sir', 'Chairman', etc.
- **Female-associated terms:** 'she', 'her', 'hers', 'herself', 'Ms.', 'Madam', 'Chairwoman', etc.

The datasets provide a comprehensive lexicon of gendered terms, making it a suitable resource for this study. The frequency of these gendered terms within each article was counted, and the dominant gender representation was determined.

4.4.2 Categorisation of Articles Based on Gender Representation

The categorisation of articles was based on the presence and frequency of gendered terms and pronouns, enabling a systematic evaluation of gender bias in university news content. Articles were classified into three distinct categories:

- **Male-Oriented (M):** Articles that predominantly featured masculine pronouns or possessive terms (e.g., 'his', 'he') as indicators of male-focused content.
- **Female-Oriented (F):** Articles that predominantly featured feminine pronouns or possessive terms (e.g., 'her', 'she') as indicators of female-focused content.
- **Neutral (N):** Articles that did not contain an identifiable predominance of male or female gender-specific terms.

This classification enabled a structured, quantitative analysis of gender representation across the dataset, allowing for a data-driven investigation into gendered language usage in academic media.

4.4.3 Statistical Analysis of Gender Representation

Once the articles were categorised, a statistical analysis was conducted to determine the proportion of male, female, and neutral-oriented articles within the dataset. This involved counting the number of articles in each category (M, F, N). Computing the proportion of male- versus female-oriented articles to identify potential gender disparities. Applying statistical tests, such as the Chi-square test, to evaluate whether the observed gender

distribution significantly deviated from an expected uniform representation. The Chi-square test is based on the following formula:

$$X^2 = \sum \frac{(O - E)^2}{E}$$

Where:

- $O =$ Observed frequency (how many articles actually fell into each category (F, M and N)).
- $E =$ Expected frequency (how many would be expected if there were no gender bias e.g., an equal split across categories).
- The sum ($\sum$) is calculated over all categories (F, M and N).

This statistical analysis provided empirical insights into the prevalence of gendered language in university news, highlighting potential overrepresentation or underrepresentation of male or female references.

Following this, sentiment analysis was conducted on both male-oriented (M) and female-oriented (F) articles to examine the tone and sentiment associated with different gender representations. Additionally, Career and Family terms analysis was performed to explore gender-related themes within professional and personal contexts. The next two sections outline the methodology and process for each analysis in detail.

4.4.4 Methodological Consideration

The lexicon-based approach was selected for its efficiency and objectivity in identifying gendered language within large textual datasets. By focusing on possessive terms and pronouns, the study effectively captured explicit gender references. However, methodological consideration must be acknowledged in term of using lexicon-based method. The lexicon-based method primarily identifies explicit gendered terms, but it does not account for implicit gender bias. For example, it does not recognise gendered occupational roles, where certain professions are disproportionately associated with one gender (e.g., 'nurse' being predominantly female-associated, or 'engineer' being predominantly male-associated).

Despite this limitation, the lexicon-based approach provided a solid foundation for investigating gender bias in academic media and laid the groundwork for further qualitative and contextual analysis.

4.4.5 Gender Distributional Frequency Analysis Results

The gender distribution analysis revealed significant disparities in the representation of male and female terms in university news articles. Out of the 5768, articles analysed,1986 (34.4%) predominantly featured male-related terms, 742 (12.9%) featured female-related terms, and the remaining 3,040 articles (52.7%) were classified as neutral, with no predominant gender terms (see Table 4.3). A Chi-square test for independence revealed that the observed distribution of gender terms in the university news articles significantly deviated from an expected uniform distribution ($\chi^2(2, N=5768) = 1354.75, p < 0.001$). This indicates a substantial imbalance in gender representation, with male-associated terms appearing far more frequently than female-associated ones, and a large proportion of articles categorised as gender-neutral. Example sentences from the dataset further illustrate the gender bias in representation:

> **M:** "The Scientific Director is Professor XX whose groundbreaking research has shown how disruption to the signalling system within a cell can trigger cancer. He has worked at some of the world's most prestigious cancer research institutions and was previously Executive Dean of the Faculty of Biological Sciences at XX."
>
> **F:** "Professor XX entered the profession in the 1980s, at a time when the overwhelming majority of physicists were men and attitudes prevailed that women were not cut out for 'hard science'. Throughout her career, she has worked to improve equality and diversity."

Class	Number of News Articles
F	742
N	3040
M	1986

Table 4.3: The number of news articles by class (N, M, and F).

Prior to proceeding to the following sections, it is important to clarify that the dataset was deliberately maintained in its original unbalanced state to accurately reflect the real-world distribution of gender representation within university news articles. Artificially balancing the data for example, by equalising the number of male- and female-oriented articles would mask the inherent biases present in the source material. The objective was not to impose artificial parity but to quantify existing disparities in how genders are portrayed, both in terms of sentiment and thematic framing (e.g., career versus family-related terms). By preserving the original imbalance, such as the significantly higher number of male-oriented articles compared to female-oriented ones, the analysis captures the authentic

landscape of media coverage and reveals whether sentiment or lexical usage differs relative to each gender's representation.

This approach aligns with established bias detection research, where the primary objective is to measure systemic imbalances rather than adjust them beforehand (Aizenberg and Van Den Hoven, 2020; Kasy and Abebe, 2021). Studies investigating gender bias in media representation emphasise the importance of preserving natural distributions to accurately quantify existing disparities rather than imposing artificial balance (Asr et al., 2021). For statistical comparisons, proportions were employed to account for differences in sample sizes, ensuring that findings reflect relative trends rather than being distorted by raw volume disparities (Johnson and Khoshgoftaar, 2019). Consequently, the decision to maintain the original distribution allows this study to transparently document bias as it naturally occurs within the dataset, providing a more faithful representation of the underlying issues in media representation.

4.5 Thematic Analysis of Gendered UoL-NA: Career and Family Terms

This section presents the thematic analysis conducted to investigate potential gender biases in the representation of career-related and family-related terms within university news articles. The primary objective was to determine whether male-oriented (M) and female-oriented (F) articles exhibited differences in thematic focus particularly in professional achievements versus family or community-oriented topics. By employing a lexicon-based approach, this analysis systematically identified career-related and family-related terms in university news articles.

A predefined lexicon-based approach was used to identify career-related and family-related terms in the dataset. The study utilised established career and family-related word lists from Dacon and Liu (2021). These lexicons were comprehensive, incorporating gender-specific equivalents (e.g., 'policeman' and 'policewoman') to ensure that gendered language was accurately captured. By systematically counting the occurrences of these terms in the dataset, the study aimed to provide a quantifiable measure of thematic focus.
The analysis explored whether male-oriented (M) and female-oriented (F) articles differed in their thematic focus by comparing the prevalence of career-related terms and family-related terms.

- **Career-related terms**: Words associated with academic and professional achievements, such as ‘research’, ‘professor’, ‘grant’ and ‘publication’.
- **Family-related terms**: Words associated with family life and caregiving, such as ‘mother’, ‘family’, ‘parenting’ and ‘childcare’.

4.5.1 Quantitative Analysis of Term Frequency Differences

To examine gender-based differences in thematic focus, the study quantified the frequency of career-related and family-related terms in male-oriented (M) and female-oriented (F) articles. Statistical analysis was conducted to determine whether the observed differences were statistically significant. Two-proportion z-tests were applied to compare the frequency of career-related terms and family-related terms in male and female articles.

The two-proportion z-test compares the proportions of a specific term category (e.g., career-related terms) between two independent groups. The formula is:

$$\mathcal{Z} = \frac{(\hat{P}_{1-}\, \hat{P}_2)}{\sqrt{\hat{P}\,(1 - \hat{P})\,(\frac{1}{n_1} + \frac{1}{n_2})}}$$

where:

- $\hat{P}_{1\ and}\, \hat{P}_2$ are the sample proportions for each group (e.g., the proportion of career-related terms in M and F articles),
- $n_1\ and\ n_2$ are the sample sizes of the two groups,
- $\hat{P}$ is the pooled proportion calculated from both groups combined.

This test determines whether any observed differences in the proportions of career-related or family-related terms are statistically significant or could have occurred by chance. By applying this method, the analysis aimed to identify systematic differences in thematic emphasis between M and F articles.

4.5.2 Methodological Considerations

The study employed a lexicon-based approach to ensure that the analysis was systematic and replicable. The following methodological considerations were accounted for:

- **Use of Established Lexicons:** The career and family-related word lists from Dacon and Liu (2021)provided a consistent and validated approach to term classification.
- **Objectivity in Term Identification:** The use of predefined word lists eliminated subjectivity, ensuring that the classification of terms was data-driven rather than interpretation-based.

- **Statistical Validation:** Applying quantitative analysis ensured that any observed differences in term usage were rigorously tested for statistical significance, preventing misinterpretation of the findings.

Despite the robustness of this approach, certain limitations must be acknowledged. While the lexicon-based approach accurately identifies explicit terms, it does not account for contextual meanings. For example, the word 'professor' may be used in a non-career-related context, but it would still be counted as a career-related term. Moreover, the study focuses on explicit mentions of career and family-related terms but does not capture implicit biases, such as assumptions about traditional gender roles (e.g., women being more involved in caregiving). Nonetheless, this approach provides a strong foundation for systematically analysing gender bias in university news reporting.

4.5.3 Thematic Analysis Results

The analysis revealed 5,695 career-related words in M articles compared to 2,176 in F articles, while family-related words appeared 4,343 times in M articles and 2,942 times in F articles. In N articles, there were 2,919 career-related words and 382 family-related words recorded. Two-proportion z-tests showed significant differences in the use of both career words ($z = 53.81$, $p = 0.000$) and family words ($z = 23.21$, $p = 0.000$) between M and F articles. Proportional analysis accounting for the different sample sizes (1,986 M articles vs 742 F articles) revealed that F articles contained nearly twice as many family-related terms per article (3.97 vs 2.19), whilst career term frequency remained comparable (2.93 vs 2.87). These results indicate that female-focused articles disproportionately emphasise family and caregiving contexts compared to male-focused articles, illustrating stereotypical gendered framing in university news coverage. Example sentences are provided below:

- **Career:**

> **M:** "Professor XX will be heading up a new Academic Unit of Palliative Care at the XX Institute of Health Sciences, a teaching and research institute within the University's School of Medicine. He will be leading research to develop and test Innovative treatments that aim at improving the care of patients who have an incurable Illness."
>
> **F:** "To investigate the importance of these 'master controller' regions in protein aggregation in living cells, the team joined forces with Dr. XX and her students, also members of the Astbury Centre in XX."

- **Family:**

> **M:** "Professor XX's successful campaign to right historical wrongs for LGBT veterans and their families demonstrates the powerful role that evidence-based research plays in influencing public policy."
>
> **F:** "But Mrs XX, now a mother and grandmother herself, is pleased work by academics such as Dr XX is shedding new light on the experiences of her father and his fellow XX."

4.6 Sentiment Analysis of Gendered News Articles

Sentiment analysis is traditionally performed using lexicon-based or supervised ML approaches, where models are trained on large datasets with sentiment-labelled text. However, Facebook BART-Large-Mnli[19] takes a different approach, leveraging NLI-based zero-shot classification to determine sentiment without requiring task-specific labelled data. This capability is made possible by its pre-training on large-scale text corpora and subsequent fine-tuning on the Multi-Genre Natural Language Inference (MultiNLI) dataset (Williams et al., 2017; Lewis, M. et al., 2019).

Instead of being explicitly trained for sentiment classification, BART-Large-Mnli reinterprets sentiment analysis as an NLI task. Given an input text (such as a university news article), the model evaluates how well it supports predefined hypothesis statements that correspond to different sentiment categories (positive, negative, and neutral). The model then assigns a probability score to each hypothesis and classifies the text based on the hypothesis with the highest probability.

This model was selected because prior research provides direct support for its use in news sentiment classification, which is closely aligned with the aim of this study. The most relevant study is that of Bucher and Martini (2024), who evaluated BART-Large-Mnli on the sentiment classification of New York Times coverage of the US economy and reported 0.85 accuracy and 0.80 macro-F1. These results suggest that the model can perform well on formal journalistic text, rather than only on reviews or social media data. Their study also compared BART-Large-Mnli with other zero-shot approaches, including GPT-3.5, GPT-4 and Claude Opus and found that BART-Large-Mnli performed strongly and consistently within that group. This is important because the present study is not intended to fine-tune a model, but rather to test a strong off-the-shelf zero-shot method. On that basis, BART-Large-Mnli is a well-justified choice, as it has already shown good performance on news text and is especially suitable where manually labelled domain-specific data for university news are limited or unavailable.

[19] https://huggingface.co/facebook/bart-large-mnli

4.6.1 Facebook BART-Large-Mnli

The Facebook BART-Large-Mnli model was trained using a two-stage process:

- **Pre-training Phase:** The model was first pre-trained as a denoising auto-encoder on large-scale unlabelled text corpora to learn general language patterns, syntax, and semantics. These corpora included:

    - **Common Crawl** – A diverse internet text dataset (Raffel et al., 2020).
    - **Wikipedia** – High-quality encyclopaedic articles (Devlin et al., 2019).
    - **BooksCorpus** – A collection of fiction and non-fiction books (Zhu et al., 2015).
    - **OpenWebText** – Curated web-based text (Radford et al., 2019).

- **Fine-Tuning Phase:** After pre-training, BART was fine-tuned on MultiNLI, a dataset designed to train models in textual entailment tasks (Williams et al., 2017). The MultiNLI dataset consists of 433,000 premise-hypothesis pairs, where the model learns to predict whether a hypothesis is entailed, contradicted, or neutral with respect to a given premise.

This fine-tuning stage enables BART-Large-Mnli to generalise to new classification tasks, including sentiment analysis, by treating sentiment as an entailment problem. Unlike traditional sentiment classifiers that rely on manually annotated datasets, Facebook BART-Large-Mnli uses an NLI-based approach. The model follows these steps when classifying sentiment:

Step 1: Input Text (Premise): The news article or sentence is treated as the premise, which is the text that the model needs to evaluate. For example, consider the following input text from a university news article:

"Professor XX was awarded a prestigious research grant."

Step 2: Hypothesis Statements for Sentiment Classification: To classify sentiment, the model is given three hypotheses, each corresponding to a sentiment category:

- Positive Sentiment Hypothesis: "This text expresses a positive sentiment."
- Negative Sentiment Hypothesis: "This text expresses a negative sentiment."
- Neutral Sentiment Hypothesis: "This text expresses a neutral sentiment."

Step 3: Entailment Prediction Using NLI: The Facebook BART-Large-Mnli model then evaluates how strongly each hypothesis is entailed (i.e., supported) by the input text. The model assigns probabilities based on three possible relationships:

1. **Entailment** → The premise supports the hypothesis.
2. **Contradiction** → The premise contradicts the hypothesis.
3. **Neutral** → The premise is unrelated to the hypothesis.

For the given text ("Professor XX was awarded a prestigious research grant."):

Hypothesis	**Probability Score**	**Classification**
"This text expresses a positive sentiment."	0.85 (85%)	Entailed (Positive)
"This text expresses a negative sentiment."	0.05 (5%)	Contradiction
"This text expresses a neutral sentiment."	0.10 (10%)	Neutral

Table 4.4: Probability scores assigned by Facebook BART-Large-Mnli for sentiment classification

Since the positive hypothesis has the highest probability, the model classifies the text as positive sentiment. The model determines the sentiment probabilities based on the following:

1. **Word Embeddings and Semantic Understanding**

Unlike lexicon-based sentiment classifiers, Facebook BART-Large-Mnli does not rely on predefined word lists. Instead, the model uses:

- **Word embeddings** → Representations of words in a high-dimensional space where similar words appear closer together (Mikolov et al., 2013).
- **Contextual relationships** → Understanding how words interact within sentences.

For instance, during pre-training, the model has likely seen sentences such as:

- "She was awarded the Nobel Prize for her groundbreaking research."
- "The university is known for its prestigious faculty and cutting-edge facilities."

From these examples, the model learns the associations between words like 'awarded', 'prestigious' and 'research grant' with positive connotations.

2. **Contextual Encoding**

Rather than classifying words individually, the transformer architecture allows BART-Large-Mnli to process entire sentences holistically. This enables it to:

- Recognise nuanced sentiment expressions**.**
- Distinguish between factual and opinion-based sentiment**.**

3. Probability Calculation

Once the premise-hypothesis pairs are processed, the model calculates probabilities for each sentiment category using the SoftMax function:

$$P(y \mid x) = softmax(f(x, h))$$

where:

- x is the input text (news article).
- h is the hypothesis.
- $f(x, h)$ is the model's score for entailment.
- $P(y \mid x)$ is the probability distribution over sentiment labels.

The hypothesis with the highest probability is selected as the final sentiment classification.

The model's performance was evaluated using a small set of manually annotated 46 articles to assess its accuracy for more details about the annotation see Chapter 3 Section 3.5.3). The subset of 46 articles was selected randomly for the sentiment analysis task, and the annotators labelled these articles as positive, negative or neutral. Since the selection was random, the class distribution was not balanced across genders based on the gender distribution analysis introduced in section 4.4.

Of the total sample, 7 articles referred to individuals identified as F, 12 referred to M, and 27 referred to N. This imbalance reflects the natural distribution within the sampled data rather than a controlled or stratified selection process. After annotation, the three independent annotators classified the articles as follows. For F articles, 3 articles were labelled positive, 2 negative, and 2 neutral. For M articles, 7 were labelled positive, 5 negative, and 0 neutral. For articles categorised as N, 9 were labelled positive, 16 negative, and 2 neutral. Overall, across all genders, annotators assigned 19 positive labels, 23 negative labels, and 4 neutral labels.

Metrics like precision, recall, and F1-score were used to evaluate the model's performance. In addition to these standard classification metrics, two-proportion z-tests were also conducted to assess whether there were statistically significant differences in sentiment distribution between article categories, particularly between M and F articles. This helped investigate potential sentiment bias across gendered content.

The two-proportion z-test compares the proportions of a specific sentiment (e.g., positive) between two independent groups. This test determines whether any observed differences in sentiment proportions are statistically significant or could have occurred by chance. By applying this method, the analysis aimed to uncover any systematic bias in how sentiment was distributed between M and F articles.

4.6.2 Sentiment Analysis Findings

The performance of the Facebook BART-Large-MNLI model was evaluated against the manually annotated ground-truth sample using precision, recall and F1-score metrics. The model achieved an overall accuracy of 85%, with high precision and recall for positive and negative sentiments. Although the neutral category returned an F1-score of 0.00, this should be interpreted cautiously because the annotated evaluation sample included only four neutral instances. The limited representation of neutral sentiment in the ground-truth data means that this result is more likely to reflect sample imbalance than a definitive weakness in the model's performance (see Table 4.5).

	Preci-sion	**Recall**	**F1**
Negative	0. 82	1.00	0.90
Neutral	0.00	0.00	0.00
Positive	0.89	0.84	0.89
Accuracy		0.85	
Macro Avg.	0.57	0.61	0.59
Weighted Avg.	0.78	0.85	0.81

Table 4.5: Evaluation metrics of Facebook BART-Large-Mnli model's sentiment

The sentiment analysis conducted using the Facebook BART-Large-Mnli model revealed that positive sentiment was present in 84.74% (1,683) of male-related articles, 85.58% (635) of female-related articles, and 83.95% (2,552) of neutral articles. Negative sentiment was found in 15.16% (301) of male-related articles, 14.42% (107) of female-related articles, and 15.95% (485) of neutral articles. Neutral sentiment was detected in 0.1% (2) of male-related articles, 0.0% (0) of female-related articles, and 0.1% (3) of neutral articles.

To investigate potential bias in sentiment distribution between gendered content, two-proportion z-tests were conducted on M and F articles. The results showed no statistically significant difference in the proportion of positive sentiment between M and F articles ($z = -0.54$, $p = 0.586$) or negative sentiment ($z = 0.48$, $p = 0.632$). Neutral sentiment could not be meaningfully compared due to near-complete absence in both categories (2 instances in M articles, 0 in F articles), reflecting the model's detection of only a very small number of neutral cases. Because neutral sentiment was identified too rarely to allow robust analysis, it was excluded from further interpretation. These findings suggest no clear evidence of sentiment bias between M and F articles, although this conclusion should be considered considering the limited number of neutral classifications. The following examples illustrate how sentiment was expressed within these two categories.

- **Positive:**

> **M**: "He has also acted as a special representative in China for the Institute of Civil Engineers, developing opportunities for British businesses both in China and further afield. Since 2017, Sir XX has chaired the Executive Group of the University's Institute for High-Speed Rail and System Integration."
>
> **F**: "XX has become a passionate ambassador against knife crime. 'I eat and sleep knife crime every day. Young people have become desensitised towards knife crime and this needs to change', she said."

- **Negative:**

> **M**: "Attack on Creative Expression in India', Dr. XX, Associate Professor in Information Technology Law, describes the jewellery company's advert, which ended up being pulled from circulation, before proceeding to discuss and analyse the nation's response to the advert, as many deemed it controversial. Dr. XX discusses the significance of cultural and religious influences of the advert as well as the wider context of legal restraint on creative expression in India."
>
> **F**: "Women who had a final diagnosis of NSTEMI had a 41 per cent greater chance of a misdiagnosis when compared with men."

- **Neutral:**

> **M**: "At XX, Professor XX is leading a five-year interdisciplinary study into the impact of climate change on the Congo peatlands."
>
> **F**: "Professor XX, Vice-Chancellor, addressed the recent pay award in an email to all staff."

A limitation of the Facebook BART-Large-MNLI model in this study relates to the classification of neutral sentiment. While the model performed well in identifying positive and negative sentiment, its performance for the neutral category was weaker, with an F1 score of 0.00. This result should, however, be interpreted cautiously, since the annotated sample used for evaluation contained only a small number of neutral cases. In the context

of university news, where language is often factual and non-evaluative, this may have reduced the model's opportunity to demonstrate neutral classification more fully. As a result, the findings indicate that the model was effective for clearer sentiment polarity, but that further validation on a more balanced sample would be needed to assess neutral sentiment more reliably.

4.7 Subjectivity in Annotation and Evaluation of DistilBERT and BART-Large-MNLI

The annotation of subjectivity bias and sentiment in the UoL-NA dataset is inherently subjective, as reflected by the relatively low agreement between annotators shown in Chapter 3, Sections 3.5.2.2 and 3.5.3.1. This makes standard accuracy metrics difficult to interpret, since the ground truth itself is not always consistent. To provide a clearer evaluation, a subset of the test set where all three annotators agreed on a single class label is considered. These cases represent the highest level of certainty, where bias or sentiment is least ambiguous.

For the subjectivity bias task, full agreement was achieved in 57 out of 110 cases (51.8%). Within this subset, DistilBERT aligned with annotators in 41 cases, giving an agreement rate of 71.9%. This indicates that when the class (i.e. bias or not bias) is clearly identifiable, the model performs reasonably well. However, the fact that nearly half of the dataset lacks consensus highlights the difficulty of the task and suggests that lower overall accuracy may partly reflect annotation subjectivity rather than model weakness.

For sentiment analysis, only 21 out of 46 cases (45.7%) had full agreement. Of these, Facebook BART-Large-MNLI matched annotator labels in 16 cases, corresponding to 76.2% agreement. As with bias detection, the model shows stronger performance on clearer examples, while disagreement among annotators limits confidence in evaluating more ambiguous cases.

Overall, both models demonstrate solid performance when assessed on high-confidence annotations. While subjectivity in labelling restricts definitive conclusions, the results suggest that the models are capable of capturing patterns that align well with human judgement where consensus exists.

4.8 Discussion

4.8.1 Insights from LLMs Classification and Bias Annotation

As this component represented the initial phase of the research, zero-shot LLMs were employed without domain-specific fine-tuning. This approach is well-justified by recent studies showing that zero-shot models achieve strong performance across various NLP tasks (Brown et al., 2020; Kojima et al., 2022; Wei et al., 2022), providing a valuable baseline for assessment before committing to resource-intensive fine-tuning procedures.

The experiments conducted in this study offer valuable insights into the comparative effectiveness of LLMs and human annotators in detecting bias within university news articles. LLMs such as GPT-3.5-turbo and Google Bard present a scalable and cost-efficient solution for large-scale classification. However, their limited accuracy significantly hinders their capacity to identify domain-specific biases, particularly those that are subtle and context-dependent. In contrast, human annotators demonstrated a stronger ability to recognise nuanced forms of bias, owing to their capacity for contextual interpretation and critical judgement attributes that current LLMs lack. This distinction highlights a fundamental trade-off: while LLMs offer scalability and affordability, they do so at the expense of reliability, whereas human annotations, although more resource-intensive, provide richer and more context-aware evaluations.

Nonetheless, the human annotation process is not without limitations. One of the main challenges lies in achieving consistent agreement across annotators. Bias in academic content often manifests in subtle ways through representational imbalance, thematic framing, or omission which can be open to interpretation. As a result, even among human annotators, variation in judgement was evident. While the overall percentage agreement stood at 52.3%, Cohen's Kappa scores revealed inconsistencies, with some annotator pairs showing only slight or even negative levels of agreement. These results suggest that annotators may have applied the guidelines differently or brought in subjective interpretations influenced by individual perspectives. Therefore, while human annotators outperform LLMs in understanding context and recognising implicit bias, their effectiveness is constrained by challenges in inter-annotator reliability.

4.8.2 Implications of Bias Findings

The findings of this study reveal critical implications regarding gender bias in university news articles, particularly in how different genders are represented and framed in institutional media. Through a multi-level analysis incorporating gender representation, sentiment classification, and thematic focus, it becomes evident that structural imbalances exist in the way academic achievements and personal narratives are distributed across genders.

The gender distribution analysis uncovered a notable disparity in representation: articles with male-oriented (M) language accounted for 34% of the dataset, whereas female-oriented (F) articles made up only 12%, and the remaining 52% were categorised as gender-neutral. A Chi-square test confirmed that this distribution was statistically significant, strongly indicating a systemic imbalance in the visibility of M versus F figures in institutional reporting. This discrepancy reflects allocational bias, wherein M receive greater representation and, by extension, more opportunities for recognition in university-affiliated media.

Moreover, the content of these articles also demonstrated a pattern of representational bias. M articles were more likely to emphasise career-related achievements, professional roles, and institutional leadership, while F articles frequently included references to family roles or community engagement. Although both male and female are described positively,

the thematic framing differs in ways that subtly reinforce traditional gender norms positioning male as career-oriented leaders and female as family-oriented contributors. Such differential framing not only reflects but may also reinforce societal stereotypes about gender roles within academia.

4.9 Ethical Considerations

One of the ethical limitations of this study relates to the use of pre-trained NLP models, including Facebook BART-Large-Mnli, DistilBERT, Google Bard, and GPT-3.5-turbo. These models are trained on large-scale, general-purpose corpora that may already embed various forms of social, cultural, or gender bias. As such, their application in a specialised academic context introduces the risk that they may reproduce or even amplify existing stereotypes, rather than identify them. Without any domain-specific fine-tuning or mitigation strategies, these models may inadvertently reflect the biases they are expected to detect.

Furthermore, human annotation is itself not free from bias. While human annotators often perform better at identifying subtle, contextualised expressions of bias, their interpretations can be highly subjective. In this study, disagreements among annotators reflected in low inter-rater agreement and negative Cohen's Kappa scores highlight the inconsistencies and personal biases that may influence how individuals classify content. The framing of academic achievements, gender roles, and sentiment can vary significantly depending on the annotator's background, understanding, and implicit assumptions.

Lastly, the reliance on a single-institution dataset introduces a degree of contextual and cultural specificity that may limit the generalisability of the findings. Bias patterns identified in this study could be shaped by institutional policies, communication strategies, or even historical gender dynamics within that university. Therefore, care must be taken not to overgeneralise the results without considering these contextual factors.

In summary, while computational methods offer valuable tools for large-scale bias detection, their use is accompanied by several ethical and practical limitations, including model bias, annotator subjectivity, institutional framing, and dataset representativeness. Acknowledging these constraints is essential to critically evaluating the findings and understanding the broader challenges of bias analysis in higher education media.

4.10 Conclusion

The main goal of this chapter was to investigate the presence and patterns of gender bias in university news articles through NLP methods. This was achieved by analysing gender representation, sentiment framing and thematic focus across a large corpus of academic news content. The chapter presented a detailed comparison between automated bias classification using PLM, LLMs and human annotation via MTurk.

The results demonstrated notable disparities in gender representation, with male-oriented articles significantly outnumbering female-oriented ones. Furthermore, thematic analysis revealed that male-focused articles were more likely to highlight professional achievements, while female-focused articles more frequently included references to family and caregiving roles. These patterns illustrate both allocational and representational bias within institutional media.

While LLMs offered a scalable approach to zero-shot bias classification, they struggled to reliably detect nuanced or domain-specific bias, often producing inconsistent results. Human annotators performed better in identifying subtle forms of bias, but their reliability was limited by subjectivity and low inter-rater agreement. Sentiment analysis using the Facebook BART-Large-MNLI model showed strong performance in identifying positive and negative sentiment, while neutral tones were less frequently detected, which may have introduced some limitation in interpretation.

Overall, this chapter demonstrates the complexity of detecting bias in higher education media and highlights the limitations of both automated and human annotation methods. It underscores the need for refined tools and practices in evaluating institutional discourse and provides a foundation for more equitable and transparent academic communication. This chapter concludes the investigation of university news, while the following chapter shifts focus to a different data and introduces an alternative analytical framework for examining bias within university reading lists.

Chapter 5

5 Authorship Diversity and Bias in University Reading Lists

5.1 Introduction

Reading lists play a central role in shaping students' academic experiences and disciplinary understanding. Often curated by module leaders, these lists act as gatekeepers to knowledge, they not only direct students to essential readings, but also implicitly communicate which voices, perspectives, and epistemologies are valued within a field (Schucan Bird and Pitman, 2020). In doing so, reading lists reflect broader institutional priorities and ideological commitments, whether consciously or not. As students engage with these texts, they are also being socialised into particular academic traditions and ways of thinking. The composition of reading lists, therefore, is not neutral as it holds significant pedagogical and political implications (Rodriguez, 2018).

Therefore, reading lists serve not only as pedagogical tools but also as a valuable textual corpus for examining bias. Their curated content captures both the thematic priorities of disciplines and the voices deemed authoritative within them. Analysing these lists through NLP techniques allows to uncover recurring patterns, dominant narratives, and notable absences, offering insight into how academic knowledge is framed and where subtle forms of bias may shape students' exposure to ideas. To date, little research has systematically applied NLP methods to reading lists, making them a novel and underexplored data source for bias detection in higher education. Parts of this chapter was published in (Bin Shiha et al., 2022) and (Shiha et al., 2025b).

5.1.1 Background and Motivation

In recent years, there has been growing scrutiny of curricular materials within higher education, especially in relation to diversity, inclusion, and decolonisation. Campaigns such as "Why is My Curriculum White?"(National Union of Students 2019) and subsequent movements in the UK and internationally have called attention to the ways in which university curricula continue to centre predominantly Western, Eurocentric perspectives whilst marginalising non-Western knowledge, histories and scholarly contributions. Scholars such as Bhambra et al. (2018); Schucan Bird and Pitman (2020), have argued that the process of decolonising the curriculum requires a critical reassessment not only of what is taught, but also of who is being read. This involves recognising the historical exclusions embedded in academic canons and reconsidering whose knowledge is legitimised in educational spaces.

Researchers have noted that reading lists not only shape what students read, but also who is given authority and legitimacy as a knowledge producer within the academic space. In particular, studies such as those by Schucan Bird and Pitman (2020) have shown that reading lists in UK higher education are often dominated by white, male, Western authors, with minimal inclusion of voices from marginalised or non-Western backgrounds. This lack of representational diversity not only perpetuates intellectual exclusion but also limits the scope of perspectives to which students are exposed.

Similarly, Ahmed, S. (2012) argues that the idea of diversity is often embraced rhetorically within institutions, while the structures that shape inclusion such as reading lists, curriculum design, and institutional culture continue to reproduce dominant norms. In this way, reading lists can act as a form of academic gatekeeping, in which particular paradigms and traditions are continuously privileged, while alternative knowledge systems remain marginalised or invisible. Such critiques have fuelled wider discussions around curriculum reform, most notably within the broader discourse on decolonising higher education. This movement has gained momentum through campaigns like "Why is My Curriculum White?" and has been supported by a growing body of academic literature (Bhambra et al., 2018), all of which call for a fundamental rethinking of how knowledge is selected, structured, and delivered in university settings.

Building on this literature, it is evident that much of the existing research on bias in reading lists has focused primarily on who is being read, that is, the demographic characteristics of the authors included in course materials. While this line of inquiry is crucial in revealing patterns of exclusion and lack of representation, it does not fully account for the content of what is being read. That is, a reading list could, in theory, feature a more diverse set of authors, yet still reproduce dominant epistemologies or frameworks that marginalise alternative perspectives. Conversely, texts authored by individuals from dominant demographic groups may still present critical or subaltern perspectives. As such, focusing exclusively on authorship risks overlooking the more subtle ways in which ideological framing, narrative tone, and thematic emphasis shape the intellectual environment of a module.

A necessary first step in examining curricular bias is identifying who is being read, a focus that aligns with much of the existing literature on reading lists and representation (Bhambra et al., 2018; Schucan Bird and Pitman, 2020). This chapter begins by following that approach, offering an analysis of authorship across two sets of university reading lists, with attention to gender, ethnicity, and geographic background. This mirrors the broader critical agenda aimed at uncovering the structural exclusions embedded in higher education and the dominance of white, Western, male voices in academic canons.

However, stopping at authorship alone risks simplifying a much more complex landscape. While authors' identities can reveal important power dynamics, they do not tell us everything about the intellectual content students engage with. Therefore, this chapter goes a step further by examining the thematic focus of selected readings. Through an analysis of reading list summaries and abstracts, it explores whether institutional or cultural context

influences not just who is cited, but also what kinds of ideas, topics, and narratives are prioritised. This shift allows for a deeper interrogation of how knowledge is framed within different academic environments, and how that framing may vary even when module topics appear aligned.

5.1.2 Objectives of the Chapter

This chapter aims to quantitatively analyse the demographic profiles of reading list authors to identify baseline patterns of representational bias. It will systematically extract and categorise the gender and ethnicity of all authors across three reading lists, specifically 212 entries from AIMESRL, 73 entries from QSIERL and 679 entries from DSRL. The chapter also seeks to assess topical diversity to determine whether institutional contexts influence curricular content focus. To achieve this, it will apply transformer-based NLP methods, including sentence embeddings with all-MiniLM-L6-v2 (Reimers and Gurevych, 2019; Wang et al., 2020), semantic matching via cosine similarity, and BERTopic (Grootendorst, 2022) topic modelling, alongside dimensionality reduction and clustering, to identify and compare thematic patterns within 40 semantically matched reading pairs from the University of Leeds and Imam Mohammad Ibn Saud Islamic University. Through these tasks, the chapter will provide measurable insights into representational bias and thematic variation, contributing directly to the overall research aim of understanding how institutional contexts shape reading list composition.

5.1.3 Contributions of The Chapter

This chapter presents the first comparative analysis of reading list demographics across Western and Middle Eastern institutions, contributing three structured datasets totalling 964 demographically annotated entries. The AIMESRL and QSIERL datasets provide unique insights into how Islamic Studies curricula differ between the UK and Saudi Arabia, while the DSRL dataset establishes baseline representational patterns within interdisciplinary Disability Studies modules.

The chapter introduces a replicable NLP pipeline combining sentence embeddings, semantic matching and topic modelling to systematically detect bias in academic curricula. This transformer-based approach uncovers latent thematic structures and representational patterns that surface-level title analysis cannot capture. Building on this methodological framework, the chapter challenges existing approaches to curriculum decolonisation by demonstrating that demographic diversity in authorship does not automatically translate to epistemic diversity. Comparative entropy analysis reveals that Leeds exhibits broader thematic range compared with Imam Mohammad ibn Saud Islamic University's more concentrated focus, emphasising that institutional context shapes both who is read and how knowledge is framed. This fulfils the methodological contribution outlined in Chapter 1 Section 1.3.4.2, designing a replicable NLP pipeline to analyse demographic and thematic diversity in university reading lists.

5.2 Reading Lists Datasets

This section draws on three datasets developed to analyse representation in reading materials:

- **AIMESRL:** Contains 212 reading list entries extracted from undergraduate modules at the University of Leeds, focusing on Arabic and Islamic Studies materials.

- **QSIERL:** Includes 73 readings assigned in modules taught at Imam Mohammad Ibn Saud Islamic University, representing Islamic educational contexts.

- **DSRL:** Comprises 679 items from eight undergraduate modules at the University of Leeds, spanning disciplines including law, literature, sociology, and education.

This chapter focuses primarily on the AIMESRL and QSIERL datasets for comprehensive cross-cultural comparison, while the DSRL dataset serves as a supplementary case study to demonstrate representation bias patterns in authorship diversity. The detailed methodology for data extraction, cleaning, standardisation, and demographic annotation of these datasets is provided in Chapter 3 section 3.3.2.

5.3 Authorship Diversity Analysis

An examination of the gender and ethnicity of authors listed in university reading materials reveals clear patterns of demographic concentration and imbalance. Across all three datasets, male authors dominate, though the ethnic composition varies significantly by institution and subject domain.

5.3.1 AIMESRL

Analysis of the 212 items in the AIMESRL dataset shows a marked dominance of male and Western authors. After excluding entries where demographic details could not be reliably determined, the gender distribution reveals that 84.91% of identifiable authors were male, and 15.09% were female. Ethnicity analysis similarly reveals a Western tilt: 44.34% of authors were identified as North American, and 36.79% as European, with only 11.32% categorised as Middle Eastern. Asian and African authors made up 6.6% and 0.94% of the dataset respectively. These results reinforce prior research suggesting that reading lists in UK higher education tend to centre white, Western, male voices (Schucan Bird and Pitman, 2020).

5.3.2 QSIERL

In contrast, the QSIERL dataset compiled from 73 readings across 27 modules reflects a regional and linguistic orientation toward the Middle East. 87.14% of authors were identified as Middle Eastern, and 97.2% were male. Only 2.8% female were. A smaller proportion of authors were classified as African 10% and Asian 2.86%. No authors were identified as North American or European. This may be due in part to the dataset being exclusively in Arabic, with many of the readings authored by scholars in Arabic-speaking countries.

5.3.3 Comparison: AIMESRL vs. QSIERL

A comparative analysis between AIMESRL and QSIERL highlights the contrast in representation (see Figure 5.1). AIMESRL's authorship is overwhelmingly Western, while QSIERL features predominantly Middle Eastern authors. Despite being within the same disciplinary domain, the two institutions curate readings that reflect their respective cultural and institutional positions. The QSIERL dataset did not include any of the same reading materials found in AIMESRL, underscoring the disconnect between Western and Arab academic canons even when discussing similar topics. Notably, there is no overlap between the two datasets: none of the reading materials assigned in QSIERL appear in AIMESRL. Furthermore, 25 of QSIERL's 73 readings are available in English, demonstrating that a substantial portion of its Middle Eastern–authored scholarship is accessible beyond Arabic-speaking contexts.

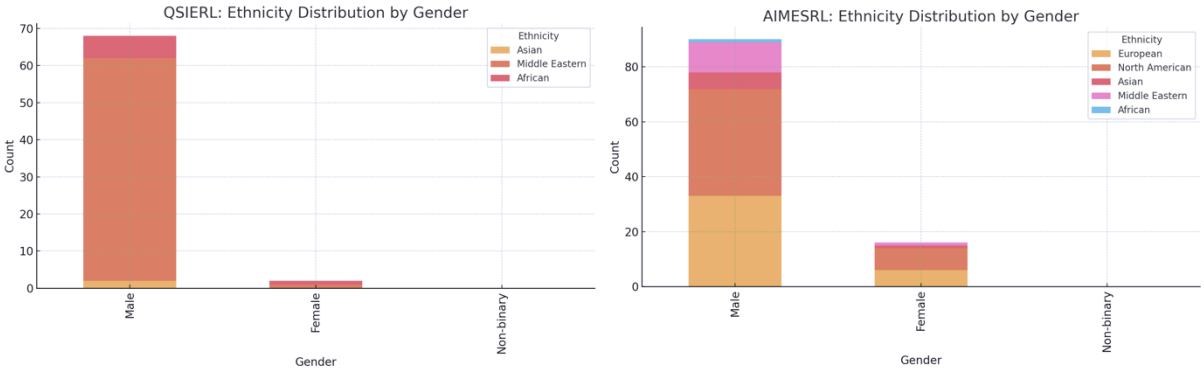

Figure 5.1: AIMESRL and QSIERL ethnicity distribution by gender

5.4 Topic-Matched Subset and Text Summary Generation (AIMESRL and QSIERL)

5.4.1 Topic-Matched Subset

To examine the thematic content of university reading materials beyond authorship, a third dataset was constructed using a controlled sample of texts from

the AIMESRL and QSIERL reading lists. This subset was designed to facilitate a comparative content analysis by focusing on readings that address similar topics across the two institutions. The objective was to explore whether comparable module themes such as Qur'anic interpretation or Islamic education are framed differently depending on institutional context.
The initial pool consisted of 211 reading list titles from AIMESRL and 71 titles from QSIERL. From these, a balanced selection of 20 readings from each dataset was produced using semantic similarity techniques to identify the most thematically aligned titles. This ensured that the comparison was not random but grounded in content-level equivalence.

Each reading list title was first embedded into a fixed-dimensional vector space using a pre-trained sentence embedding model. Specifically, the model used was "all-MiniLM-L6-v2", part of the "SentenceTransformers" framework (Reimers and Gurevych, 2019) and based on the "MiniLM" architecture developed by Wang et al. (2020). Sentence-BERT models such as this are fine-tuned using contrastive and triplet loss on semantic similarity tasks, enabling them to capture rich contextual and lexical features beyond word-level matching.

Formally, let each reading list title be represented by a sentence ss, which is mapped to an embedding vector $V_s \in \mathbb{R}^d$, where dd is the dimension of the embedding (e.g. $d = 384$ for MiniLM). These embeddings are constructed such that semantically similar sentences occupy nearby positions in the vector space.

To compare two titles $s_1 \; and \; s_2$ with corresponding embeddings $v_1 \; and \; v_2$, cosine similarity was used:

$$\cos \_sim(v_1, v_2) = \frac{v_1 * v_2}{[\![v_1]\!][\![v_2]\!]}$$

This metric outputs a value between -1 and 1, where higher values indicate greater semantic similarity. A similarity matrix was generated by computing cosine scores between every possible title pair from AIMESRL and QSIERL. The top-scoring pairs were manually reviewed to ensure topical relevance, resulting in a final set of 40 thematically matched readings 20 from each institution. The use of only 40 readings was intentional for this proof-of-concept study, prioritising methodological demonstration over statistical generalisability. While this sample size limits robust institutional comparison, it provides sufficient data to demonstrate how transformer-based methods can reveal thematic and representational structures within academic curricula.

This embedding and matching process is visually summarised in Figure 5.2, which illustrates how each title is transformed into a dense vector and compared with titles from the other dataset using cosine similarity.

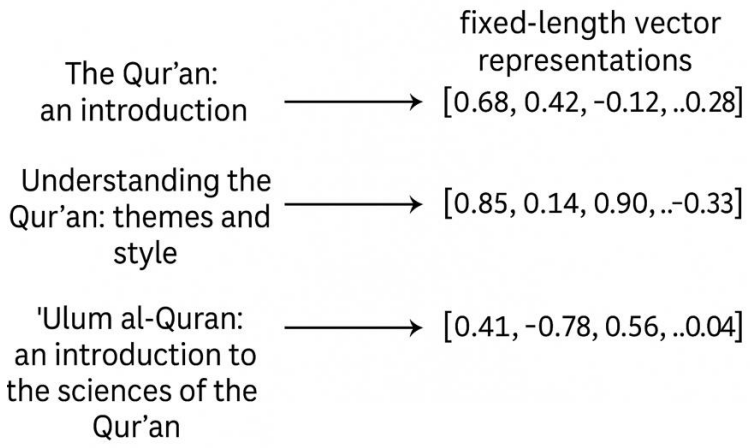

Figure 5.2: Converting reading list titles into fixed-length vector representations using sentence embeddings

Each reading list title is encoded using a transformer-based model (e.g. MiniLM (Wang et al., 2020)), producing a vector that represents its meaning. Cosine similarity is then used to compare vectors and identify thematically aligned readings across AIMESRL and QSIERL.

5.4.2 Text Summary Generation

To facilitate content-level analysis of the matched reading list items, it was necessary to obtain a consistent textual representation of each reading's core themes. While some texts included abstracts or summaries in their original bibliographic sources, others particularly older or less formally published works lacked any descriptive content. In many cases, the available descriptions varied significantly in length, detail, and quality.

To address this, a LLM, specifically OpenAI's GPT-4.o, was used to either retrieve or generate a brief abstract or summary for each selected reading. Where a reliable existing summary could be found online through web search (e.g. publisher websites, Google Books, or academic repositories), that summary was used. In cases where no such description was available or where existing text was too limited or inconsistent, GPT-4.o was prompted to generate a short summary based on the title and any available contextual clues (e.g. author name or field).

This approach provided a standardised form of representation across all 40 readings in the topic-matched dataset, ensuring that each item could be analysed in a comparable way. Using GPT-4.o for summary generation allowed for greater consistency in both length and style, reducing bias introduced by variable abstract formats and enabling a more coherent input for subsequent topic modelling.

While AI-generated summaries are not perfect substitutes for reading full texts, they serve as a practical and scalable approximation of thematic content particularly useful for comparative corpus analysis at scale.

It is important to note that no translation was required for the subset of 20 texts drawn from QSIERL, as all selected materials were already available in both Arabic and English. Therefore, the composition of the 40 texts was not influenced by translation constraints in this case. However, some of the reading list summaries from both reading list QSIERL and AIMESRL were difficult to obtain, and GPT-4.o was used to construct summaries for those texts. Specifically, ten of the QSIERL summaries and seven of the AIMESRL summaries were generated using GPT-4.o in cases where existing summaries were not readily available.

The generation process utilised iterative refinement of standardised prompts, with the primary instruction "Generate a concise academic-style summary (2-3 sentences) of this reading based on its title and known academic context" supplemented by domain-specific guidance emphasising historical and interpretive focus for classical texts. Model parameters included a temperature setting of 0.2 to reduce output variability whilst maintaining coherence, with a maximum constraint of 150 tokens to ensure concise, focused summaries. All GPT-4.o generated summaries underwent manual review, with revisions made where necessary to ensure factual accuracy and thematic relevance following established academic writing conventions.

To ensure the reliability and consistency of GPT-4.o generated content, a validation framework was implemented comprising four analytical dimensions, following recent methodological advances in AI-generated content evaluation (Liu, Y. et al., 2023; Wang, Z.P. et al., 2024).

1. Semantic Similarity Analysis

Cosine similarity analysis was employed to assess the semantic alignment between title and summary embeddings, generated using the all-MiniLM-L6-v2 sentence-transformer model. This approach demonstrated that the summaries consistently captured the thematic content of their corresponding titles, with a mean similarity score of 0.743, indicating strong overall alignment. The standard deviation (SD) of 0.111 suggested relatively low variability across the dataset, while the observed similarity scores ranged from 0.494 (minimum alignment) to 0.904 (maximum alignment). Although cosine similarity has recognised limitations in certain contexts (Zhou et al., 2022; Steck et al., 2024), it remains a widely adopted metric for validating semantic consistency in text embeddings, particularly when applied to models such as sentence-transformers that normalise vector representations (Reimers and Gurevych, 2019).

2. Term preservation analysis

Term preservation analysis examined the retention of key terms from titles to summaries, revealing excellent content coverage. On average, 86.6% of substantive title terms defined as words with at least four characters, excluding common function words such as articles, prepositions, and conjunctions were present in the corresponding summaries. The SD of 0.286 reflected moderate variation in coverage across the dataset. This high rate of lexical overlap demonstrates strong thematic continuity between source titles and generated abstracts. In addition, the generated summaries exhibited consistent structural characteristics, remaining within appropriate ranges for concise academic writing and demonstrating suitable paragraph development. Cross-institutional comparisons further indicated minimal variation, suggesting the absence of systematic generation bias across different institutional sources.

3. Readability assessment:

Readability was evaluated using the Flesch Reading Ease (FRE) test, a widely adopted metric developed by Flesch (1948). The FRE score is calculated using the formula:

$$FRE = 206.835 - (1.015 \times ASL) - (84.6 \times ASW)$$

Where ASL is average sentence length (words per sentence) and ASW is average syllables per word. Scores range from 0 to 100, with higher values indicating easier readability. For example, scores between 60 and 70 correspond to "plain English" typical of everyday newspaper articles, while scores below 30 are characteristic of very difficult texts commonly seen in academic or highly technical writing (Elmira, 2025).
In this study, FRE scores for the GPT-4.o generated summaries averaged 12.9 (SD = 11.2), positioning them firmly in the "very difficult" category. This outcome reflects the expected academic register of abstracts in Islamic Studies, with their technical vocabulary and complex sentence structures.

4. Summary characteristics

An analysis of structural characteristics demonstrated that the generated summaries exhibited strong consistency with the conventions of concise academic abstracts. The average summary length was 142.5 words (SD = 60.3), falling within the target range for effective academic summaries, which are typically expected to remain under 150–200 words (Hartley, J., 2003). This consistency indicates that the model successfully produced abstracts of an appropriate scope for scholarly communication. The generated summaries averaged 5.3 sentences per summary (SD = 2.5), reflecting well-structured paragraph development that balances conciseness with sufficient elaboration of key ideas. Cross-institutional comparisons revealed minimal variation in summary length, with Leeds summaries averaging 130.0 words and Imam summaries averaging 155.7 words, indicating the absence of systematic generation bias and demonstrating that the model produced consistently structured summaries across different institutional contexts.

To illustrate the validation process, a representative example from the corpus demonstrates the quality and consistency of the generated summaries. For instance, the title "Religious Policies of the Caliphs from Al-Mutawakkil to Al-Muqtadir" was summarised as follows: "Christopher Melchert's article examines the religious policies of the Abbasid caliphs during this period through their judicial appointments. Melchert challenges the notion that al-Mutawakkil fully reestablished traditionalism, noting that his judicial appointments suggest only limited support for that tendency. His successors al-Muntaṣir, al-Mustaʿīn, and al-Muʿtazz did not pursue substantially different policies…" At 156 words, this summary clearly preserves key historical details, maintains a scholarly analytical tone, and stays closely aligned with the original title, showing that the validation framework effectively produces high-quality, consistent summaries.

Overall, these results show that the GPT-4.o–generated summaries were both accurate in meaning and well-structured, following standard conventions for academic abstracts.

5.5 Topic Modelling and Semantic Matching

To examine the thematic structure of the matched summaries from AIMESRL and QSIERL, this study employed BERTopic (Grootendorst, 2022), a flexible topic modelling framework that integrates transformer-based sentence embeddings, dimensionality reduction, density-based clustering, and keyword extraction. This method is particularly well-suited for short documents such as summaries and abstracts, where traditional topic modelling methods like LDA (Blei et al., 2003) often struggle with coherence.

The process begins by embedding each summary $d_i \in D$, where D is the corpus of document summaries, into a high-dimensional vector space using the same pre-trained sentence embedding model all-MiniLM-L6-v2 that was previously applied during the title matching stage. This model, part of the "SentenceTransformers" framework (Reimers and Gurevych, 2019), is based on the "MiniLM" architecture developed by Wang et al. (2020). The model encodes each summary into a 384-dimensional vector that captures rich semantic and contextual features, providing a consistent embedding approach across both stages of analysis. This results in an embedding matrix:

$$\mathbf{E} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n], \qquad \mathbf{v}_i \in \mathbb{R}^{384}$$

Where $\mathbf{v}_i$ represents the embedding of summary d_i in a 384-dimensional semantic space. These embeddings form the input for the clustering process.

To reduce dimensionality while preserving both local and global structure, the embeddings were projected into a lower-dimensional space using Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018). Let $\mathbf{E}_{\mathbf{UMAP}} \in \mathbb{R}^{n \times r}$ denote the reduced matrix, where $r \in \{2,5\}$ depending on the desired resolution or visualisation setup. UMAP was used with its default parameters (n_neighbors=15, min_dist=0.1), which balance local semantic fidelity with global separability.

The resulting vectors were then clustered using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) (Campello et al., 2013), which identifies dense regions in the data and assigns clusters without requiring a predefined number of topics. Each cluster T_k represents a latent topic within the dataset, and all documents assigned to a cluster are treated as belonging to the same thematic category. HDBSCAN was applied using its default configuration, including a min_cluster_size of 10, which is suitable for short-text datasets with modest sample sizes. These defaults were retained to ensure consistency and reproducibility, avoiding the risk of overfitting through manual tuning.

To extract interpretable labels for each topic, BERTopic applies class-based Term Frequency Inverse Document Frequency (c-TF-IDF)(Sparck Jones, 1972). This technique consolidates all documents within a topic into a single class and then calculates a modified term weighting for each word t in topic k using the following formula:

$$\mathrm{c-TF-IDF}_{t,k} = \frac{f_{t,k}}{\sum_{t'} f_{t',k}} * \log\left(\frac{N}{n_t}\right)$$

Where:

- $f_{t,k}$ is the frequency of term t in topic k,
- $\sum_{t'} f_{t',k}$ is the total frequency of all terms in topic k,
- n_t is the number of topics in which term t appears,
- N is the total number of topics.

This weighting emphasises terms that are distinctive to a given topic while penalising terms that are common across all clusters. The top-scoring keywords under this metric form the basis of the topic labels. This behaviour is particularly valuable in thematic analysis, as it ensures that each topic is defined by terms that are distinctive rather than generic. Words that frequently occur across all topics such as "Islam," "book," or "study" tend to be less informative in distinguishing one cluster from another. By assigning higher weight to words that are common within a single cluster but infrequent elsewhere, class-based TF-IDF improves the semantic coherence and interpretability of the resulting topic labels, making it easier to identify the unique conceptual focus of each group.

To assess the thematic diversity of the readings assigned by each institution, the distribution of documents across topics was analysed using Shannon entropy (Shannon, 1948), calculated as:

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i$$

Where:

- p_i is the proportion of documents assigned to topic i,
- n is the total number of topics.

Higher entropy values indicate a more evenly distributed thematic spread, while lower values suggest that documents are concentrated around fewer dominant topics.

The overall analytical pipeline is summarised visually in Figure 5.3, which illustrates the step-by-step process applied in this study.

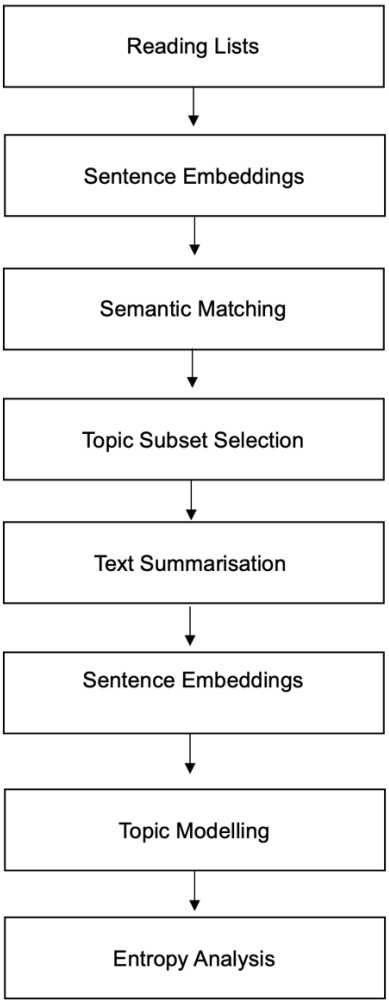

Figure 5.3: Workflow for transformer-based thematic and demographic reading list analysis

5.5.1 BERTopic Implementation and Parameter Configuration

5.5.1.1 Preprocessing Pipeline

Prior to topic modelling, all summary texts underwent systematic preprocessing to ensure consistency and remove noise. The preprocessing pipeline comprised five sequential stages:

1. **Case normalisation**: All text was converted to lowercase to ensure consistent token matching across the corpus.
2. **Character filtering**: Non-alphabetic characters were removed using the regular expression r'[^a-zA-Z\s]', retaining only letters and whitespace to eliminate potential noise from punctuation and special characters.
3. **Stop word removal**: A comprehensive stop word list combining NLTK English stop words and scikit-learn's ENGLISH_STOP_WORDS was applied to eliminate common function words that might obscure meaningful semantic distinctions.
4. **Tokenisation**: Texts were split using whitespace-based tokenisation for processing efficiency whilst maintaining semantic integrity.
5. **Text reconstruction**: Cleaned tokens were rejoined with single spaces to create the final processed summaries suitable for embedding generation.

This preprocessing approach prioritised simplicity and reproducibility whilst maintaining the semantic content necessary for meaningful clustering analysis.

5.5.1.2 Model Configuration and Parameter Selection

BERTopic was implemented using carefully selected components optimised for short academic texts. The configuration employed:

- **Sentence embeddings**: The all-MiniLM-L6-v2 model, producing 384-dimensional dense vector representations specifically designed for semantic similarity tasks.
- **Dimensionality reduction**: UMAP with default parameters including n_neighbors=15 (balancing local and global structure preservation), min_dist=0.1 (allowing moderate compression whilst maintaining cluster separation), and metric='cosine' (appropriate for high-dimensional text embeddings).
- **Clustering algorithm**: HDBSCAN with default configuration, including min_cluster_size=10 (though effectively reduced due to the modest corpus size) and min_samples=None (automatically determined based on min_cluster_size).
- **Topic representation**: c-TF-IDF with default n-gram range (1,1) to highlight distinctive terms within each identified cluster.

The decision to employ default parameters was based on BERTopic's established performance on similar short-text academic corpora and the exploratory nature of this proof-

of-concept study. Previous research has demonstrated the effectiveness of BERTopic's default configuration on academic texts, including studies on NLP abstracts (Samsir et al., 2023) and scientific article analysis (Chagnon et al., 2024). Whilst parameter optimisation might yield marginal improvements, the default configuration provided stable and interpretable results suitable for methodological demonstration.

5.6 Topic-Matched Subset and Topic Analysis Results

5.6.1 Topic-Matched Subset

Following the semantic similarity process, a final set of 40 thematically aligned readings was identified 20 from AIMESRL and 20 from QSIERL. This matching was conducted using transformer-based sentence embeddings and cosine similarity, ensuring that the selection reflects strong content-level equivalence rather than superficial lexical overlap.
Some examples from the top-matched pairs illustrate the thematic coherence achieved through this process:

- "The Qur'an: An Introduction" (AIMESRL) and "Introduction to the Study of the Qur'an" (QSIERL), both introductory texts aimed at familiarising students with the structure, themes, and interpretive traditions of the Qur'an.
- "Understanding the Qur'an: Themes and Style" and "The Meanings of the Qur'an", which offer literary and thematic explorations of Qur'anic content.
- "ʿUlūm al-Qurʾān: An Introduction to the Sciences of the Qurʾān" and "The Perfect Guide to the Sciences of the Qur'an", both presenting foundational concepts in Qur'anic sciences.
- "Muʿāwiya Ibn Abī Sufyān: From Arabia to Empire" and "Muḥarrar al-Wajīz – Tafsīr Ibn ʿAṭiyyah", pairing a historical-political biography with a classical tafsīr that engages similar periods and figures.

The cosine similarity scores for these matches ranged from 0.96 down to 0.58, with manual inspection confirming that the top pairs displayed high topical alignment in Qur'anic studies, Islamic historiography, and interpretive methodology. These aligned pairs form a content-rich basis for cross-institutional comparison and subsequent analysis.

To facilitate this comparison, a brief summary was generated for each of the 40 matched readings. These summaries were either retrieved from reliable bibliographic sources or generated using GPT-4.o, ensuring consistent and thematically representative abstracts across the dataset. Although the final set includes only 40 readings, this constrained sample size was a deliberate design choice to prioritise semantic alignment and interpretive depth across institutions. The study functions as a proof-of-concept, not a generalisable statistical analysis.

5.6.2 Topic Analysis

To explore the thematic structure of the 40 matched readings, topic modelling was applied using the BERTopic framework. The model used all-MiniLM-L6-v2 from the sentence-transformers library as its embedding model, which was passed to BERTopic via the embedding model argument during initialisation. Dimensionality reduction was handled by UMAP, and clustering was performed using HDBSCAN, the default unsupervised density-based algorithm integrated into BERTopic.

BERTopic successfully identified two distinct clusters:

- **Topic 0: Qur'anic Tafsir and Interpretive Methodologies:** This was the largest cluster, containing 24 documents. It was defined by keywords such as "quran," "quranic," "tafsir," "ibn," "work," and "islamic". Representative documents in this cluster included classical exegesis works such as "Jāmiʿ al-Bayān ʿan Taʾwīl Āy al-Qurʾān", commonly known as Tafsīr al-Ṭabarī.
- **Topic 1: Early Islamic History:** The second cluster contained 16 documents, with top terms like "islamic," "historical," "early," "history," and "sources". Representative readings here included texts like "Roger Mervyn Savory's Introduction to Islamic Civilisation", which survey foundational historical developments in the Muslim world.

To assess the breadth of thematic coverage within the reading lists of the two institutions, Shannon entropy was calculated based on the distribution of topics identified through BERTopic. Shannon entropy is a measure of diversity that increases when topics are more evenly represented and decreases when a small number of topics dominate the dataset. It provides a quantitative lens through which to evaluate the extent to which each institution's readings are topically balanced or concentrated.

The results reveal a clear disparity in topical diversity between the two institutions. The reading list from Leeds yielded a Shannon entropy score of 0.791, indicating a moderately diverse topic distribution. This suggests that the selected texts are spread relatively evenly across different thematic clusters, allowing for broader topical exposure in the curriculum. The University of Leeds appears to exhibit greater thematic diversity, with readings spanning Qur'anic exegesis as well as historiography, ethics, and reformist thought, suggesting a more interdisciplinary approach to Islamic Studies.

In contrast, the reading list from Imam showed a significantly lower entropy score of 0.336, reflecting a more concentrated topical focus. This means that a majority of Imam's texts fall into a limited number of topics, suggesting a narrower thematic range with greater emphasis on classical tafsīr and Qur'anic sciences. Such a distribution could imply a more specialised or traditional curricular emphasis, potentially at the expense of exposing students to a wider array of subjects or interpretive frameworks. This concentration reflects a pedagogical focus grounded in traditional scholarly frameworks.

These findings offer an important empirical dimension to the discussion of curricular representation. They suggest that students at Leeds may encounter a more varied set of discourses within Islamic and Middle Eastern studies, whereas students at Imam may engage more deeply with a smaller number of themes. This disparity in topical diversity raises further questions about educational scope, institutional priorities, and potential gaps in scholarly exposure.

However, whilst these entropy values suggest potential differences in thematic breadth between the two institutions, the small sample size limits the statistical significance of this comparison. These preliminary findings are consistent with different pedagogical orientations, yet the use of entropy as an analytic tool demonstrates a potentially scalable method for evaluating thematic distribution, though validation with larger datasets would be needed to confirm its effectiveness for curriculum auditing.

5.6.3 Clustering Stability and Validation

To address concerns regarding the reliability and robustness of the topic modelling approach, a comprehensive validation framework was implemented comprising multiple complementary analyses.

5.6.3.1 Bootstrap Stability Analysis

Clustering stability was assessed through bootstrap resampling to evaluate the consistency of topic assignments across different data samples. The validation protocol involved 50 bootstrap iterations using 80% random sampling without replacement, with each bootstrap sample independently processed through the complete BERTopic pipeline. Stability was measured using the Adjusted Rand Index (ARI) calculated for overlapping documents between bootstrap iterations, alongside assessment of topic count variability through standard deviation across iterations.

Results indicated moderate clustering stability (ARI: 0.485), which, whilst below the threshold typically considered indicative of high stability (>0.5), represents reasonable consistency for exploratory analysis of a small, specialised corpus. The topic count remained remarkably stable (1.4 ± 0.9), suggesting robust identification of the underlying thematic structure despite the modest sample size.

5.6.3.2 Parameter Sensitivity Analysis

Robustness was further evaluated through systematic parameter sensitivity analysis, focusing on the key clustering parameter min_cluster_size, which determines the minimum number of documents required to form a cluster in HDBSCAN. This parameter directly influences the trade-off between cluster granularity and noise detection, with smaller values typically producing more fine-grained topics and larger values creating broader, more inclusive clusters (Campello et al., 2013).

Testing was conducted using min_cluster_size values of 3, 5, 8, 10, and 15. Across all parameter settings, the model consistently identified two primary topics, with zero documents assigned to noise categories. Topic size distributions remained stable across parameter variations, with the larger cluster consistently containing 60-65% of documents and the smaller cluster 35-40%, suggesting robust thematic partitioning regardless of clustering sensitivity.

The absence of noise assignments is particularly noteworthy, as HDBSCAN typically labels marginal or ambiguous documents as outliers under more heterogeneous conditions (McInnes et al., 2017). This pattern suggests that the corpus exhibits substantial internal thematic coherence, with clear semantic boundaries between the identified clusters. The consistency across parameter settings indicates that the observed two-topic structure reflects genuine thematic distinctions rather than artifacts of parameterisation.

However, this robustness should be interpreted within the context of the corpus characteristics. The specialised, domain-focused nature of the Islamic Studies texts may contribute to the observed stability, as documents share common terminological and conceptual frameworks. In larger, more heterogeneous datasets, parameter choice typically exerts stronger influence on cluster structure (Hahsler et al., 2019), and the present findings may not generalise beyond similarly focused academic corpora.

5.6.3.3 Topic Coherence Evaluation

Topic quality was assessed through coherence analysis, which evaluates whether topic-defining words co-occur within the original summaries. Following established approaches for coherence measurement in topic modelling (Röder et al., 2015), a pairwise co-occurrence measure was employed, calculating the proportion of word pairs from each topic's top 10 terms that appeared together within individual summaries. Coherence scores were subsequently averaged across all word pairs within each topic.

This approach was selected over traditional coherence metrics such as UMass or CV coherence due to the short and domain-specific nature of the corpus. Conventional reference corpora (e.g., Wikipedia) may provide misleading baselines for specialised academic content, as external corpora often lack the specific terminology and conceptual frameworks present in Islamic Studies literature (Newman et al., 2010; Lau et al., 2014). The internal coherence approach ensures that evaluation reflects the actual co-occurrence patterns within the target domain rather than general language usage.

The analysis yielded exceptionally high coherence scores: an overall average of 0.989, with Topic 0 achieving perfect coherence (1.000) and Topic 1 scoring 0.978. These values indicate that topic-defining words consistently co-occur within documents, supporting the interpretability and validity of the identified thematic clusters. The near-perfect coherence likely reflects the specialised, tightly focused nature of the corpus, where domain-specific terminology is deployed consistently across texts. For context, coherence scores above 0.7 are typically considered indicative of well-formed topics in academic literature (Mimno

et al., 2011), making these results exceptionally robust for the specialised corpus under examination.

5.6.3.4 Embedding Quality and Cluster Separation

Finally, the quality of the underlying semantic representations was examined through embedding-based evaluation. Two complementary measures were employed: silhouette analysis, which compares intra-cluster cohesion against inter-cluster separation, and centroid distance analysis, which quantifies the Euclidean separation between topic centres in embedding space. Together, these metrics provide a balance between local (document-level) and global (topic-level) perspectives on cluster quality.

The silhouette score of 0.137, while below the conventional >0.3 threshold associated with well-separated clusters, should be interpreted in context. Low silhouette scores are expected in domains where topics overlap semantically, as is often the case in Islamic Studies curricula where thematic boundaries are fluid rather than rigid. Importantly, the high centroid distance (0.814) indicates that the two topics are nonetheless well differentiated at the global level. Taken together, these results suggest that while the thematic distinctions are subtle at the document level, they nonetheless correspond to genuine and interpretable orientations at the corpus level.

5.6.3.5 Topic Distribution by Authors

An analysis of author demographics across the reading lists of both Imam Mohammad ibn Saud Islamic University and the University of Leeds reveal a predominant representation of male authors, particularly those of Middle Eastern origin. This pattern aligns with historical trends in Islamic scholarship, where male scholars have traditionally held prominent roles in the interpretation and transmission of religious texts (Mernissi, 1991; Alwani, 2013; Ahmed, L., 2021). Studies have highlighted that while women have contributed significantly to Islamic scholarship, their roles have often been underrepresented in mainstream academic discourse (Mernissi, 1991; Alwani, 2013; al-Nadwi, 2021).

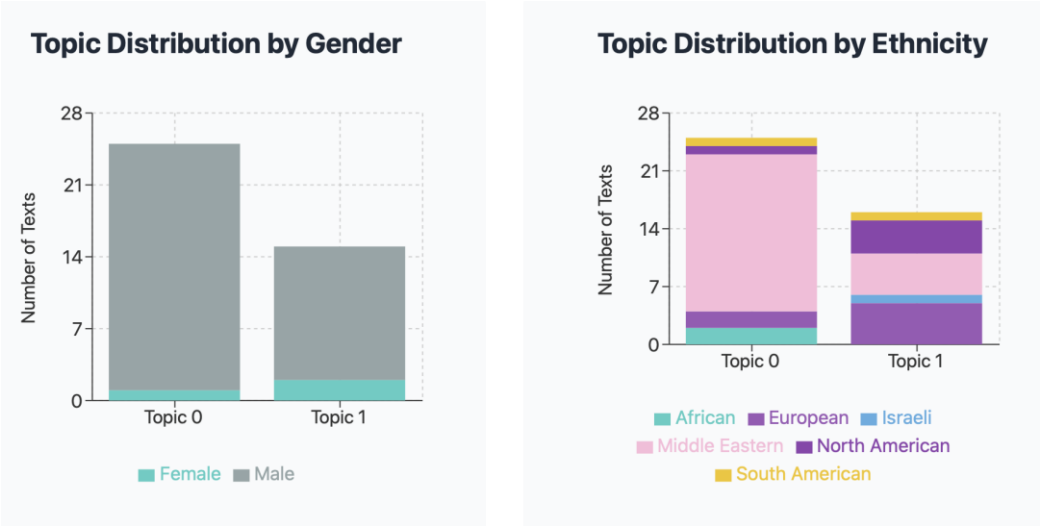

Figure 5.4: Author demographics across topic clusters: gender and ethnicity representation.

The ethnic composition of authors also varies between the two institutions. Imam University's reading list predominantly features Middle Eastern scholars, reflecting its focus on traditional Islamic sciences and its geographical and cultural context. In contrast, the University of Leeds showcases a more diverse range of authors, including those from Western backgrounds, indicative of its broader academic approach to Islamic and Middle Eastern studies.

Figure 5.4 offers a visual summary of how author backgrounds are distributed across the identified thematic clusters. While it is important to avoid assuming that authors of a particular gender or ethnic background necessarily share uniform perspectives or produce homogenous content, the distribution patterns observed here suggest that certain topics may be more commonly associated with particular demographic groups. For example, texts authored by Middle Eastern males are predominantly concentrated in Topic 0, which is primarily focused on Qur'anic sciences and tafsīr. In contrast, Topic 1, encompassing themes such as historiography, gender, ethics, and reform, includes a more varied set of authors, including female scholars and those from diverse ethnic and geographic backgrounds.

These findings echo concerns raised in prior research about the demographic imbalances in academic reading lists, particularly within UK higher education. Previous studies have found that white, Eurocentric/Western, and male authors tend to dominate the prescribed literature in British universities, contributing to the underrepresentation of alternative voices and perspectives (Phull et al., 2019; Schucan Bird and Pitman, 2020). Although demographic identity should not be equated with intellectual stance, the alignment of certain themes with particular author profiles may reflect deeper institutional and disciplinary structures that shape knowledge production and curricular design. In this context, demographic analysis not only reveals representational disparities but also offers a

lens through which to critically examine the thematic contours of what is taught, and whose voices are made central within academic knowledge frameworks.

5.6.3.6 Topical Focus Across Institutions

While there is some overlap in topical content between the two universities, the distribution of materials shows that each institution places greater emphasis on different thematic clusters. Topic modelling of the reading lists further underscores these distinctions (see Figure 5.5).

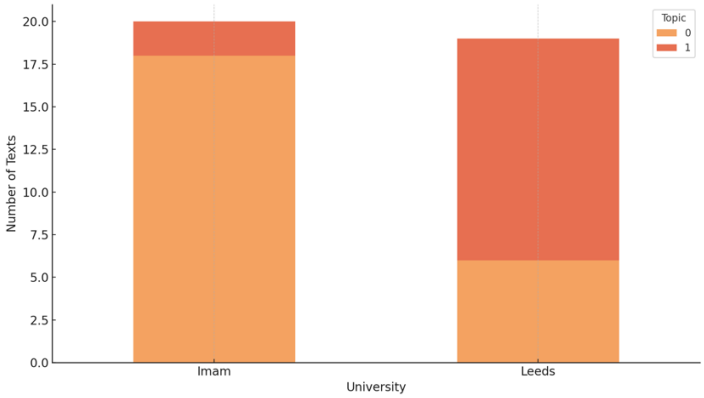

Figure 5.5: Topic distribution by university reading lists

At Imam Mohammad Ibn Saud Islamic University, the materials are heavily concentrated in Topic 0, which encompasses Qur'anic sciences and tafsīr. This concentration reflects a curriculum deeply embedded in classical Islamic scholarship. The study of tafsīr (exegesis of the Qur'an) has long been regarded as a foundational pillar of the Islamic intellectual tradition. It is not only one of the earliest disciplines developed by Muslim scholars but also one that continues to shape the interpretation of religious doctrine, law, and ethics (McAuliffe, 1991; El-Mesawi, 2005; Görke and Pink, 2014). Qur'anic sciences (ʿUlūm al-Qurʾān), which include disciplines such as the collection of the Qur'an, reasons for revelation (asbāb al-nuzūl), abrogation (naskh), and modes of recitation (qirāʾāt), form the methodological and epistemological basis through which Islamic texts are understood and applied (Wansbrough and Rippin, 1977; Von Denffer, 1983; Abdul-Rahim, 2011; Taftahi and Nazari, 2017).

Conversely, the reading list from the University of Leeds demonstrates broader thematic coverage across several distinct clusters. According to the topic modelling analysis, the materials are distributed primarily within Topic 1, which encompasses a diverse range of areas including Islamic historiography, gender and ethics, political reform, modern intellectual thought, and comparative theology. Specific readings explore the historical evolution of Islamic legal systems, critical approaches to hadith, Islam and feminism, Muslim responses to modernity, and the intellectual contributions of reformist and contemporary

thinkers. This indicates a curriculum designed to provide students with thematic variety and interdisciplinary engagement, reflecting the pluralism of scholarly voices and approaches within the field of Islamic and Middle Eastern studies.

While the titles selected for both institutions were semantically aligned at the surface level through title matching, the topic modelling of their summaries reveals significant differences in the underlying content. This divergence in topic distribution despite the presence of title-level similarity highlights the limitations of assuming equivalence based on superficial semantic overlap. The University of Leeds curriculum, with its broader thematic range, appears to prioritise coverage across multiple domains, exposing students to varied methodological frameworks and critical lenses. By contrast, the more concentrated topical structure at Imam Mohammad Ibn Saud Islamic University, focused primarily on tafsīr and Qur'anic sciences, suggests a curriculum centred on the in-depth study of foundational Islamic texts and classical epistemologies.

These differences in content coverage underscore the varying pedagogical priorities of the two institutions. Importantly, they also offer a valuable entry point into investigating how curricular design can reflect or perpetuate epistemic bias. Analysing not only who is included on reading lists, but also what topics are addressed and how, contributes to the broader project of decolonising the curriculum. This approach foregrounds the structural dynamics through which certain kinds of knowledge "textual, historical, reformist, or otherwise" are privileged, shaping students' exposure to the intellectual diversity within Islamic studies.

5.7 Discussion

The findings presented in this chapter offer critical insights into how knowledge is framed within university curricula through the composition of reading lists. While both institutions examined operate within the broad disciplinary umbrella of Islamic and Middle Eastern Studies, the analysis demonstrates that the content and structure of their reading lists reflect markedly different epistemological orientations.

Topic modelling and entropy analysis suggest that students are not merely being exposed to different authors, but also to fundamentally different thematic priorities. Imam's reading list is deeply rooted in classical Islamic sciences, particularly Qur'anic tafsīr and interpretive methodologies, revealing a concentrated and traditional scholarly focus. In contrast, the Leeds curriculum presents a more thematically diverse range of readings, covering historiography, gender, ethics, reform, and modern intellectual thought. While both institutions share some overlap in core topics especially those that fall within Topic 0 the breadth of coverage at Leeds and the depth of focus at Imam suggest different pedagogical orientations one emphasising thematic variety, the other prioritising continuity within core disciplinary traditions.

Importantly, these content patterns intersect with, but do not always mirror, authorial demographics. The University of Leeds includes a more diverse set of authors in terms

of ethnicity and gender, whereas Imam's readings are almost entirely written by Middle Eastern male scholars. Yet, the analysis cautions against assuming that a more demographically diverse authorship automatically results in epistemic diversity. Prior studies (e.g. Phull et al. (2019); Schucan Bird and Pitman (2020); Bin Shiha et al. (2022)) have shown that curricula in Western universities often remain shaped by dominant Eurocentric frameworks even when featuring authors from underrepresented backgrounds. Conversely, scholars from traditionally dominant groups may still present critical or subaltern perspectives. This complexity reinforces the need to move beyond surface-level inclusion toward a deeper interrogation of the ideas, narratives, and paradigms being conveyed.

The findings also speak directly to the broader discourse on curriculum decolonisation. Students across many UK institutions have called for more inclusive, reflective curricula that challenge hegemonic narratives and represent a plurality of worldviews (Charles, 2019). The evidence here supports the notion that decolonising the curriculum is not simply a matter of diversifying authorship, but of critically examining how knowledge is organised, legitimised, and communicated. What is taught, its themes, language, and interpretive frameworks is as important as who is being read.

In sum, this study underscores that reading lists are not neutral repositories of information; they are curated expressions of institutional priorities and academic worldviews. They frame the scope of what students learn, the perspectives they encounter, and the voices that are normalised or marginalised within their academic formation. Understanding both the demographic composition and thematic structure of reading materials thus offers a richer, more nuanced picture of curricular representation one that is essential for any meaningful movement toward more inclusive and critically engaged education.

5.8 Limitations

While this study provides valuable insights into the thematic and demographic structures of university reading lists, several limitations should be acknowledged. First, the analysis was based primarily on summaries and abstracts rather than the full content of the texts. Although summaries offer a useful approximation of a work's focus, they may not capture the full depth, nuance, or argumentative complexity of the original material. As such, conclusions drawn from topic modelling should be interpreted as indicative rather than definitive representations of textual content.

Second, the use of LLM-generated summaries helped standardise and complete the dataset, particularly in cases where official abstracts were unavailable. However, these AI-generated summaries may lack the stylistic distinctiveness, rhetorical emphasis, and disciplinary tone of the original authors. While they contributed to a more uniform input for topic modelling, they inevitably introduced a degree of abstraction from the authentic voices and intentions of the texts.

Third, the sample size for content analysis was relatively small, comprising 40 matched texts across two institutions. While these were carefully selected to ensure balance and

thematic relevance, the limited scale restricts the breadth of the conclusions. Additionally, the study focused on reading lists from only two academic departments Islamic Studies which, although meaningful for a comparative case study, do not capture the full diversity of curricular design across the broader spectrum of higher education. As a result, the findings should be understood as specific to the institutional and disciplinary contexts examined here, rather than generalised across all academic programmes.

Finally, several methodological limitations merit consideration. The clustering approach identifies nuanced rather than stark thematic distinctions, with validation metrics reflecting subtle but genuine differences in curricular emphasis between institutional contexts. The moderate clustering stability (ARI: 0.485) and silhouette scores (0.137), whilst below conventional thresholds, are appropriate for exploratory analysis of specialised academic corpora where thematic boundaries are inherently less distinct than in heterogeneous text collections.

The reliance on default BERTopic parameters, whilst ensuring reproducibility and avoiding overfitting, may not represent optimal configurations for this specific corpus. However, parameter sensitivity analysis demonstrated consistent topic identification across different settings, suggesting that the core findings are robust to methodological variations within reasonable bounds.

5.9 DSRL: Interdisciplinary Author Diversity Analysis

The third dataset in this study, consisting of 679 author entries from DSRL modules at The University of Leeds. As discussed in Chapter 3, section 3.3.2.4, the DSRL dataset was treated differently due to its reliance on edited volumes. Chapter authors were used for demographic coding and multi-author works were retained as single entries. The dataset was structured to capture title, chapter title, author information and demographic labels, with ethnicity coded using broad regional categories: North American, European, Australian, Asian, African and South American.

This data reveals a relatively balanced gender distribution among identifiable authors, alongside substantial gaps in ethnicity data (see Figure 5.6). Approximately 34.5% of authors were female, 32.5% male, and 0.3% non-binary, while 32.7% had unspecified gender suggesting that where gender could be determined, representation was broadly comparable across male and female authors. The ethnicity breakdown, however, is far more limited: 81.1% of authors had unclear ethnicity, with only 10.9% identified as North American and 6.3% as European. Other categories (Australian, African, Asian, South American) represented less than 1% each. Among identifiable authors, the underrepresentation of authors from BAME backgrounds is notable: African and Asian authors collectively made up only 1% of the dataset, compared to 17.2% for North American and European authors combined.

Figure 5.6 illustrates the ethnicity distribution by gender among this identifiable subset around 128 authors, 18.9% of the total. Among identifiable authors, North American and

European scholars collectively dominated at 17.2%, while African and Asian authors combined represented less than 1%. It is important to acknowledge that the high proportion which is 81.1% of unidentifiable authors is a significant methodological limitation the true ethnicity distribution across the full DSRL dataset cannot be determined with confidence. This gap likely reflects the challenges of inferring demographic information from publicly available sources alone, particularly for authors who maintain a limited online presence or whose names do not clearly indicate geographic or ethnic origin. As a result, the findings presented here should be interpreted as indicative of patterns among identifiable authors only, rather than as a definitive representation of the full dataset.

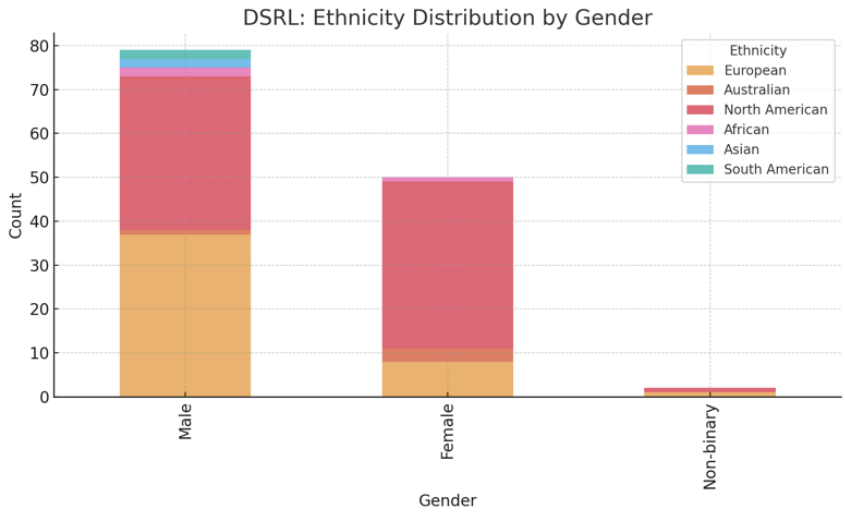

Figure 5.6: DSRL ethnicity distribution by gender

The interdisciplinary nature of disability studies spanning modules from disability law and inclusive education to medical and humanities presents opportunities to engage with diverse global perspectives on disability. However, the data suggests this breadth across legal, educational, medical and cultural frameworks has not translated into demographic diversity among cited authors, particularly given the limited ethnic diversity identified even within the 18.9% of authors where classification was possible. Modules such as disability rights and the international policy context and inclusion would benefit from incorporating voices from the global south and diverse ethnic backgrounds, where experiences of disability intersect with different cultural and socioeconomic contexts. These patterns suggest that even fields explicitly committed to challenging exclusion require ongoing attention to ensure their reading lists reflect the diverse perspectives central to their foundational values.

5.10 Conclusion

This chapter has explored how university reading lists can reveal patterns of representation and topical emphasis, offering insight into the ways knowledge is framed within academic institutions. By combining topic modelling with demographic analysis, the study identified evidence of bias in both authorship and content. Male authors, particularly those of Middle Eastern background, dominated the reading list from Imam Mohammad Ibn Saud Islamic University, with a strong focus on classical Islamic sciences. The University of Leeds, while still male-dominated, exhibited greater authorial diversity and a broader thematic range, particularly in areas related to historiography, ethics, and reformist discourse. Although some overlap existed between the two institutions, particularly around Qur'anic studies, the differences in topical distribution and demographic alignment suggest divergent epistemological orientations. Methodologically, the study demonstrates how natural language processing specifically topic modelling can support critical curriculum auditing by revealing latent patterns in textual data. This approach provides a scalable and interpretable means of analysing not only what is being taught but how knowledge is structured across institutions. Importantly, it shows that surface-level similarities, such as shared titles, do not necessarily equate to shared content or perspectives. Reading lists, often overlooked as mere administrative tools, thus emerge here as both a valuable data source and a lens into the epistemic values of academic departments. They reflect broader institutional priorities, disciplinary traditions, and the politics of inclusion and exclusion. As such, they are a powerful site for both analysis and intervention in ongoing efforts to build more inclusive, reflexive, and critically engaged curricula.

Overall, this work lays the groundwork for more sophisticated bias detection systems capable of operating in sensitive and high-stakes domains such as education. It emphasises the value of combining automated approaches with expert human insight and contributes both a practical methodology and a valuable resource to support future research in NLP, fairness, and inclusive education. While the analytical method demonstrated here shows promise, its adaptability to other disciplines remain to be confirmed through further empirical research, and no general claims are made beyond the Islamic Studies context. This chapter concludes the investigation of university reading lists, while the following chapter turns to a different dataset and introduces a new analytical framework for examining bias within higher education learning resources.

Chapter 6

6 Fine-Tuned LLM for Bias Analysis in Online Higher Education Learning Resources in Humanities

6.1 Introduction

Part of the research conducted and discussed in this chapter has been published in the following work (Shiha et al., 2025a).

6.1.1 Background and Motivation

While higher education learning resources are often perceived as objective or neutral, research increasingly challenges this assumption, revealing that they may contain subtle yet influential forms of bias. Rathbun and Turner (2012) argue that academic development practices and resources are shaped by underlying structures of power and authority, casting doubt on the feasibility of true neutrality in educational content. Similarly, Bhopal et al. (2018) highlight how institutional practices within higher education can marginalise ethnic minority students and staff, with academic materials contributing to these systemic inequalities. Mahmud and Gagnon (2023) further demonstrate that racial disparities in educational outcomes in the UK are partly reinforced by biased curricula, suggesting that the content of university-level texts plays a role in shaping exclusionary academic cultures.

Despite growing interest in the application of AI and NLP for bias detection, most existing studies have focused on public discourse such as media and social networks where bias tends to be more explicit and easier to identify. Comparatively little attention has been paid to academic materials, where bias is often embedded in complex linguistic and contextual cues, making it less perceptible to traditional detection techniques. This oversight represents a significant gap in current research, especially given the formative influence of educational texts on learners' perspectives.

Addressing this gap, this chapter investigates the application of LLMs, to detect both explicit and implicit biases within online higher educational content. It places particular emphasis on identifying bias related to ethnicity and gender, drawing on a curated dataset of academic texts.

6.1.2 Objectives of the Chapter

This chapter aims to achieve the following specific objectives:

To fine-tune a GPT-4o-mini model for domain-specific automated bias detection in OER-H. Moreover, to analyse model performance across different bias categories, identifying specific strengths and limitations in detecting explicit versus implicit bias forms within academic discourse.

To evaluate the fine-tuned model's performance against established baselines, including zero-shot LLMs demonstrating measurable improvements in bias detection capabilities.

To assess cross-domain generalisability by applying the humanities-trained model to UoL-DSPDS-LT, identifying transferability patterns and domain-specific bias characteristics.

6.1.3 Contributions of The Chapter

This chapter makes contributions to the field of automated bias detection in educational contexts, addressing the contributions outlined in Chapter 1. The chapter fulfils the LLM bias investigation goals (Section 1.3.3.2) by fine-tuning a domain-specific bias detection model tailored to Humanities disciplines. Additionally, it provides empirical evidence of model performance across academic disciplines by conducting a cross-domain transfer experiment: applying the humanities-trained model to STEM content to evaluate its capacity to detect bias across distinct disciplinary contexts. This cross-domain transferability analysis reveals important domain-specific bias patterns and limitations that have not been previously examined in the literature. This addresses the contribution outlined in Chapter 1, Section 1.3.3.3.

6.2 Dataset Collection and Annotation Process

The primary dataset in this investigation is OER-H, which includes text segments extracted from resources focusing on Humanities disciplines. Each segment ranges from 150-250 words and was obtained through keyword searches targeting gender and ethnic bias terminology (e.g., 'gender roles', 'racial discrimination', 'colonial narratives'). The dataset underwent comprehensive annotation. Each text segment was classified into one of three bias classes ('Bias', 'Not Bias', 'Potential Bias') with mandatory justifications explaining the classification decision (i.e., rationale for bias or non-bias determination). Additionally, text segments classified as 'Bias' or 'Potential Bias' were further categorised into one or more specific bias categories: representation bias, stereotype bias, and linguistic bias. The annotation process also required identification of specific keywords or phrases contributing to the bias classification, which encompassed both overtly biased language (e.g., 'black reputation', 'yellow coward' (Stony Brook University, no date)) and more subtle terms that may initiate bias through implication or context. Following annotation and data balancing to address class imbalance, the final dataset consisted of 108 instances with equal representation across all three bias classes (36 instances each). For all the data collection and annotation details see Chapter 3 section 3.3.3 and 3.6. The examples below

present annotated entries from the dataset, illustrating the structure and key components of the annotations.

Text Segment 1:

- **Text Segment:**
  "The controlling majority of the Federal Government, under various pretences and disguises, has so administered the same as to exclude the citizens of the Southern States… [text continues]"
- **Bias Class:** Bias
- **Bias Category:** Representation
- **Bias Keywords/Phrases:**
  - "odious and unconstitutional restrictions"
  - "destroying the institutions of Texas and her sister slaveholding States"
  - "infamous combinations of incendiaries and outlaws"
  - "imbecility of the Federal Government"
- **Justification:**
  This text reflects strong regional and political bias, portraying Northern States and the Federal Government negatively. The use of highly charged and inflammatory language presents a one-sided Southern perspective typical of secessionist rhetoric.
- **Confidence Score:** 100

Text Segment 2:

- **Text Segment:**
  "Essay ask each student to write a two page essay that answers the lessons guiding question…"
- **Bias Class:** Not Bias
- **Bias Category:** –
- **Bias Keywords/Phrases:** –
- **Justification:**
  This is an instructional prompt with no biased language, providing an open-ended academic task.
- **Confidence Score:** 100

Text Segment 3:

- **Text Segment:**
  "…from the 1860s through the 1870s the American frontier was filled with Indian wars and skirmishes… force the Indians to give up their lands…"
- **Bias Class:** Potential Bias
- **Bias Category:** Linguistic, Representation
- **Bias Keywords/Phrases:**
  - "Indian wars and skirmishes"
  - "Indian uprisings"
  - "force the Indians to give up their lands"
- **Justification:**
  The terminology frames Native American resistance in a way that downplays the violence and coercion involved in land dispossession, reflecting a biased historical narrative.
- **Confidence Score:** 100

A secondary domain transfer validation dataset is UoL-DSPDS-LT. This dataset serves to test cross-domain model performance, examining whether bias detection methods developed on humanities materials can effectively generalise to technical disciplines traditionally assumed to maintain objectivity. This dataset has not undergone any annotation process and was gathered solely to evaluate model performance across different academic domains. For detailed data collection see Chapter 3 Sections 3.3.3.

6.3 Data Augmentation

To address the relatively small size of OER-H (i.e., 108 entries, 34,248 words after balancing the annotated data), we employed contextual data augmentation, generating additional examples for each bias category to enrich the training data. This approach allowed the model to encounter a broader spectrum of biased and unbiased scenarios, ultimately improving its ability to generalise across diverse educational content.

OpenAI's GPT-4o was used to create new educational text segments designed to reflect each of the bias classes: 'Bias', 'Not Bias', and 'Potential Bias'. These additional examples were generated with a token length limit of 200 tokens per entry.

The model was given access to some of the original annotated data to learn its structure and produce examples that align with the existing dataset, which focuses on a range of Humanities disciplines. The goal was to maintain the balance of the dataset even after augmentation by ensuring that newly generated examples accurately reflected the bias classes and categories assigned to the original data. The augmentation prompt used was:

> "Generate an educational text segment suitable for higher education, focusing on themes of historical events, societal issues, or autobiographical narratives. The content should reflect complex topics such as social movements, inequality, or personal accounts of historical experiences. Ensure the segment contains the bias category and should be classified as bias class. The content should provoke critical thought, be representative of higher education discourse, and align with the themes found in history, social sciences, and humanities education.
> For each generated example, you must also provide additional metadata, including:
>
> **Justification** – Provide an explanation of why the text was classified as either 'Bias', 'Potential Bias' or 'Not Bias'. The justification should clearly explain the reasoning behind the classification, citing specific elements of language, representation, or stereotype that influenced the decision.
>
> **Bias Keywords or Phrases** – Identify words or phrases that contribute to the bias (for 'Bias' and 'Potential Bias' examples).
>
> **Bias Category** - for 'Bias' and 'Potential Bias' examples further categories them with one or more of the following:
> Stereotype Bias: Generalising roles, attributes, or behaviours to a specific gender, ethnic group, or social class.
> Representation Bias: Under-representation or limited portrayals of certain genders, ethnicities, or social groups, leading to exclusion or marginalisation.
> Linguistic Bias: Use of language that favours or promotes one group over others, or a tone that reflects implicit bias.
>
> **Confidence Score (0-100%)** – Measure how confident you are in the accuracy and correctness of the classification details provided (Bias Class, Bias Category, Bias Keywords, and Justification). If the classification is straightforward and aligns strongly with clear indicators (e.g., an obvious stereotype or a fully neutral statement), confidence should be high (90-100%). If there is some ambiguity in classification (e.g., overlapping bias categories, subtle wording, or room for debate), confidence should be moderate (70-85%). If the classification required significant interpretation or could reasonably be disputed, confidence should be lower (≥50%)".

The alignment between the generated examples and OER-H dataset demonstrates consistency in bias classification across key themes, including historical narratives, gender roles,

civil rights, literature, and immigrant integration. The generated examples successfully capture the complexity of higher educational discourse, preserving the depth needed for detecting subtle biases in educational materials.

One notable area of alignment is the representation of Western intellectual dominance. The generated example states that "the greatest philosophers and scientists have primarily emerged from Western traditions, showcasing the intellectual superiority of these cultures", reinforcing a Eurocentric bias. This aligns with the original dataset's framing of Malcolm X's legacy as "often interpreted through a Western-centric lens", illustrating how historical narratives can marginalise non-Western intellectual traditions.
Similarly, gender stereotyping appears in both datasets. The generated text claims that "women have always been more suited for caregiving roles, which is why they excel in nursing and teaching", reinforcing traditional gender roles. The corresponding original text discusses how "women's push for broader roles outside their homes faced societal resistance", reflecting historical biases that restricted their professional opportunities.

The framing of civil rights movements is another point of alignment. The generated example states that "the struggle for civil rights was a commendable effort, but at times, it pushed too aggressively, leading to unnecessary disruptions". A similar concern appears in the original dataset: "if King admits that breaking laws in order to protest is necessary, does that justify all forms of civil disobedience?" Both examples reflect bias in language and tone, framing activism as potentially excessive.
The underrepresentation of women in literature is also evident. The generated claim that "literature has historically been dominated by male writers because their works hold higher intellectual and philosophical value" aligns with the original dataset's discussion of "why women's contributions to literature were historically overlooked". Both texts highlight bias in representation within academic discourse.

Finally, bias in discussions of immigrant integration appears in the generated statement: "hardworking individuals tend to integrate well, whereas those who fail to adopt local customs often struggle to succeed". The original dataset presents a related framing, discussing "assimilation patterns of immigrant communities and their economic success". Both suggest that immigrant success is often measured by conformity to dominant cultural norms.

These examples demonstrate strong contextual alignment between the generated and original datasets, preserving thematic integrity while expanding coverage. This ensures that the dataset effectively captures bias in higher education texts while maintaining consistency in classification. A total of 75 additional examples were incorporated into the dataset, with 25 examples per bias class. While this does not represent a significant increase in dataset size, the primary objective was to expand the dataset without overshadowing the original entries. The new examples were carefully designed based on patterns observed in the original data, ensuring alignment with existing classifications. Since the original dataset had already been annotated by human experts, these additions aimed to preserve its integrity while improving the model's exposure to diverse instances of bias. As a result, the dataset (i.e., OER-H) now

consists of 183 entries, striking a balance between augmentation and maintaining the original data distribution.

To ensure data integrity and prevent leakage, all augmented examples were exclusively placed in the training set, while the test set contained only original human-annotated segments. This separation guarantees that no synthetic data influenced model evaluation, maintaining the independence required for valid performance assessment. The train/test split was carefully audited to confirm no overlap or similarity between augmented training examples and original test examples.

6.4 Fine-Tuning Setup

The fine-tuning process for bias detection was conducted using GPT-4o mini[20]. This model was selected due to its efficient performance on smaller datasets, making it well-suited for OER-H dataset. Although there is limited research directly comparing GPT-4o mini to larger models, existing evidence suggests that smaller models can achieve competitive performance without requiring extensive fine-tuning. This characteristic makes them more sample-efficient and reduces the risk of overfitting, whereas larger models often rely on large quantities of training data to generalise effectively (Eldan and Li, 2023; de Carvalho et al., 2024). Additionally, its lighter computational footprint allows for faster training and fine-tuning, making it a practical choice for iterative experimentation and evaluation.

6.4.1 Dataset Conversion to Chat Form

OER-H dataset was converted into a chat format to align with the requirements for fine-tuning GPT-4o mini model. Each text segment was structured as a 'user' message, while the corresponding bias classification, bias category, bias keywords, justification, and confidence score were formatted as the 'assistant' message. This design enabled the model to learn from a simulated conversational task, mirroring real-time bias detection.

To guide the model's responses, a system message provided context and instructions, ensuring a consistent and structured approach to classification:

> "You are a bias-detection expert reviewing higher educational materials. Your task is to analyse the given text and determine:
>
> 1. Bias Decision: Classify the text as one of the following:
> Bias: The text contains elements of stereotyping, misrepresentation, or biased language.
> Not Bias: The text is neutral, with no discernible bias.

[20] https://platform.openai.com/docs/models/gpt-4o-mini

> Potential Bias: The bias is less explicit but may be inferred depending on context or reader perception.
>
> 2. If biased, categorise it under one or more of the following categories:
>
> Stereotype: Generalising roles, attributes, or behaviours to a specific group.
>
> Representation: Under-representation or limited portrayals of certain groups, leading to exclusion or marginalisation.
>
> Linguistic: Use of language that favours or promotes one group over others.
>
> 3. Highlight specific biased words or phrases found in the text.
>
> 4. Provide a clear explanation supporting your classification.
>
> 5. Confidence Score (0-100%): Indicate how confident you are in the accuracy of the classification, based on how well the bias decision, category, keywords, and justification align with the text segment. Higher confidence (90-100%) should be assigned to cases with clear classification and unambiguous indicators, while lower confidence (≥50%) should be used when classification details are more subjective or open to interpretation".

Each classification followed a structured JSON format to ensure standardised outputs for model training. Below is an example of a text segment converted into the chat format for model training:

User Input (Text Segment):

> "I think they should be called mathematicians, not computers. I believe this because I have many black women in my college who are just as smart as white women and black or white men".

Assistant Response (GPT-4o-mini classification output):

> **Bias class:** Bias,
> **Bias category:** Stereotype, Representation,
> **Bias keywords:** "just as smart as white women and black or white men", "should be called mathematicians, not computers",
> **Justification:** The statement 'just as smart as white women and black or white men' reinforces stereotypes by implying that intelligence in black women needs to be compared to that of white women or men, positioning white individuals as the default standard for intelligence. Additionally, the phrase 'they should be called mathematicians, not computers' could be interpreted as undervaluing the historical contributions of black women who

> worked as 'human computers' in scientific fields without fully acknowledging their achievements.,
> **Confidence score:** 66.67%

6.4.2 Training Parameters and Costs

Fine-tuning was conducted using GPT-4o mini, with multiple runs performed to optimise model performance. The cost of fine-tuning GPT-4o mini is $3 (£2.37) per 1 million tokens, and the total cost across all fine-tuning runs amounted to $0.74 (£0.58). Inference costs for GPT-4o-mini were negligible, totalling less than $0.01 (<£0.01) for both input and output tokens combined.

The model was fine-tuned using the OpenAI API with the following hyperparameters: learning rate multiplier of 2.0, batch size of 1, and 3 training epochs. These parameters were selected based on OpenAI's recommendations[21] for small datasets to prevent overfitting while ensuring adequate learning.

6.5 Baseline Models

To provide a comprehensive evaluation of model performance, several baseline models were employed for comparison:

- **Zero-Shot GPT-4o mini baseline:** The base (non-fine-tuned) GPT-4o-mini model was evaluated using zero-shot prompting to directly assess the impact of fine-tuning.

- **Zero-Shot Claude-4 baselines:** Two state-of-the-art zero-shot LLMs (Claude-4[22]'s Sonnet and Opus) were prompted to classify the test set without any fine-tuning, using the same structured output as the fine-tuned model.

For these baselines, the prompt utilised is precisely the same as that used for the fine-tuned GPT-4o mini model.

This prompt instructs the model to act as a bias-detection expert and to analyse each text segment, classifying it as 'Bias', 'Not Bias', or 'Potential Bias'. Where bias is detected, the model must further indicate one or more relevant bias categories 'Stereotype', 'Representation', or 'Linguistic' and provide a justification. By employing an identical prompt structure across all LLM-based approaches, direct and fair comparison is ensured.

[21] https://platform.openai.com/docs/guides/fine-tuning

[22] https://docs.anthropic.com/en/docs/about-claude/models/overview

The zero-shot LLMs selects the most appropriate label for each text segment by estimating the likelihood of each possible class, given the input and the prompt. Formally, the model's prediction $\hat{y}$ is given by:

$$\hat{y} = \arg \max_{y \in \text{ Bias, Not}\sim\text{Bias, Potential}\sim\text{Bias}} P_{\text{LLM}}\,(y\sim|\sim x,\ \text{prompt})$$

In this formulation, the model evaluates the probability $P_{\text{LLM}}(y\sim|\sim x,\ \text{prompt})$ for each possible class y , conditioned on the input text x and the prompt. The predicted label $\hat{y}$ is the one with the highest probability. This framework encapsulates the decision-making process of zero-shot LLMs, whereby the final output is determined by whichever class the model deems most probable in light of both the input and the explicit task instructions.

6.6 Results and Analysis

The fine-tuned GPT-4o mini model was trained and evaluated on a dataset of 183 text segments, divided into a training set (80%, 146 entries) and a test set (20%, 37 entries). The data were balanced across the three bias classes (Bias, Not Bias, and Potential Bias). Performance was evaluated along two dimensions: bias class prediction and bias category identification. The model classified each text segment as 'Bias', 'Potential Bias', or 'Not Bias' and, where relevant, assigned one or more bias categories ('Stereotype', 'Representation' and 'Linguistic'). Three accuracy metrics were reported (see Table 6.1). Strict accuracy requires that the model correctly predicts both the bias class and all associated bias categories for a prediction to be counted as correct. For example, if the ground truth is 'Bias' with categories 'Stereotype' and 'Representation' and the model predicts 'Bias' with only 'Stereotype', the prediction is considered incorrect under strict accuracy, as not all categories are correctly identified. Relaxed accuracy requires that the model correctly predicts the bias class and at least one of the associated bias categories for a prediction to be counted as correct. In the same example, predicting 'Bias' with only 'Stereotype' would be considered correct under relaxed accuracy. Class accuracy measures the proportion of correct predictions for the bias class alone, regardless of category assignment. This less stringent metric reflects whether the model correctly identified the overall class as 'Bias', 'Potential Bias', or 'Not Bias'. Moreover, all precision, recall, and F1-score metrics reported use the relaxed accuracy definition. To contextualise the fine-tuned model's performance, a series of baseline models were also evaluated using the same test set: Zero-Shot LLMs Baselines (GPT-4o mini, Claude-4's Sonnet and Opus).

Model/Baseline	Strict Accuracy	Relaxed Accuracy	Class Accuracy	Relaxed Precision	Relaxed Recall	Relaxed F1-Score
Fine-Tuned GPT-4o mini	**56.76%**	**81.1%**	**81.1%**	**85.2%**	**81.6%**	**81.2%**
Zero-Shot GPT-4o mini	43.24%	67.6%	67.6%	73.0%	68.6%	66.6%
Zero-Shot Claude 4 Opus	37.84%	66.7%	72.2%	74.0%	73.9%	71.5%
Zero-Shot Claude 4 Sonnet	43.24%	64.9%	70.3%	74.6%	71.4%	68.6%

Table 6.1: Performance Comparison of Fine-Tuned and Baseline Models on Bias Detection

The fine-tuned GPT-4o mini achieved the highest performance across all metrics, with a relaxed accuracy and class accuracy of 81.1%, strict precision of 85.2%, strict recall of 81.6%, and a strict F1-score of 81.2%. This demonstrates that fine-tuning not only improved the model's ability to correctly identify bias but also enhanced its balance between precision and recall. The zero-shot GPT-4o mini baseline performed notably lower, with a strict accuracy of 67.6% and F1-score of 66.6%, highlighting the benefit of fine-tuning even when starting with a strong model. Among the other zero-shot models, Claude 4 Opus achieved a slightly higher class accuracy (72.2%) and F1-score (71.5%) compared to Claude 4 Sonnet (70.3% and 68.6%, respectively).

Across the model outputs, Linguistics emerged as the most frequently identified bias category overall, making it the most consistently predicted category across the fine-tuned and zero-shot systems. This pattern is especially visible in the fine-tuned GPT-4o mini, Claude 4 Opus, and Claude 4 Sonnet outputs, whereas zero-shot GPT-4o mini showed a slightly stronger tendency to predict Stereotyping. When considered alongside the strict accuracy results, this suggests that the models were generally better at recognising certain category signals particularly linguistic bias than at recovering the complete set of annotated categories for each instance. This is consistent with the lower strict accuracy scores for the zero-shot models, where predictions often appear to capture only part of the full bias label set. The fine-tuned GPT-4o mini nevertheless achieved the highest strict accuracy at 56.76%, compared with all the for zero-shot baselines, indicating that fine-tuning improved the model's ability to predict the full bias structure more reliably. The section 6.6.2 below provides a more detailed breakdown of bias-category predictions for each model.

6.6.1 Analysis of Bias Classes

Table 6.2 presents the per-class performance of the fine-tuned GPT-4o mini model alongside the zero-shot models bias detection across the categories 'Bias', 'Not Bias', and 'Potential Bias'. Perfect precision/recall scores should be interpreted cautiously given the small test set size (37 examples).

The results highlight that all four models perform best at identifying explicit 'Bias' cases, with high precision and recall scores (ranging from 0.91 to 0.96 F1 for this class). This indicates strong confidence in predicting explicit bias when it is present, with very few false positives or negatives. For the 'Not Bias' class, the fine-tuned model demonstrates strong recall (1.00) and improved precision (0.67) compared to its zero-shot counterpart (0.52). The zero-shot models show lower precision for this class (0.52-0.59), reflecting that some biased texts are incorrectly labelled as neutral.

The most challenging class for all models is 'Potential Bias'. This category shows the lowest recall and F1-scores, with the zero-shot GPT-4o mini particularly struggling (0.31 recall, 0.42 F1). Fine-tuning provides substantial improvement here, boosting recall from 0.31 to 0.62 and F1 from 0.42 to 0.73. However, even the fine-tuned model finds this category challenging, indicating the inherent difficulty of distinguishing between ambiguous, implied, or less explicit forms of bias and truly neutral content. Consistent with previous findings, 'Potential Bias' is most frequently misclassified as 'Not Bias', demonstrating that even advanced LLMs often overlook nuanced or implicit bias in academic texts. The comparison between fine-tuned and zero-shot GPT-4o mini clearly demonstrates that fine-tuning helps models better recognise subtle bias indicators that would otherwise be missed. The fine-tuned GPT-4o mini significantly outperforms its zero-shot counterpart, demonstrating clear benefits from fine-tuning. Compared to the zero-shot GPT-4o mini baseline, fine-tuning improved F1-scores across all classes, particularly for 'Potential Bias' and 'Not Bias'. All models face difficulty with subtle, ambiguous cases, but fine-tuning provides substantial improvements in handling nuanced bias detection.

Model	Class	Precision	Recall	F1-score
Fine-Tuned GPT-4o mini	**Bias**	1	0.83	0.91
	Not Bias	0.67	1	0.80
	Potential Bias	0.88	0.62	0.73
Zero-Shot GPT-4o mini	**Bias**	1	0.83	0.91
	Not Bias	0.52	0.92	0.67
	Potential Bias	0.67	0.31	0.42
Zero-Shot Claude 4 Opus	**Bias**	0.92	1	0.96
	Not Bias	0.59	0.83	0.69
	Potential Bias	0.71	0.38	0.50
Zero-Shot Claude 4 Sonnet	**Bias**	1	0.83	0.91
	Not Bias	0.57	1	0.73
	Potential Bias	0.67	0.31	0.42

Table 6.2: Per-Class precision, recall, and F1-score for fine-tuned and zero-Shot LLMs on bias classification

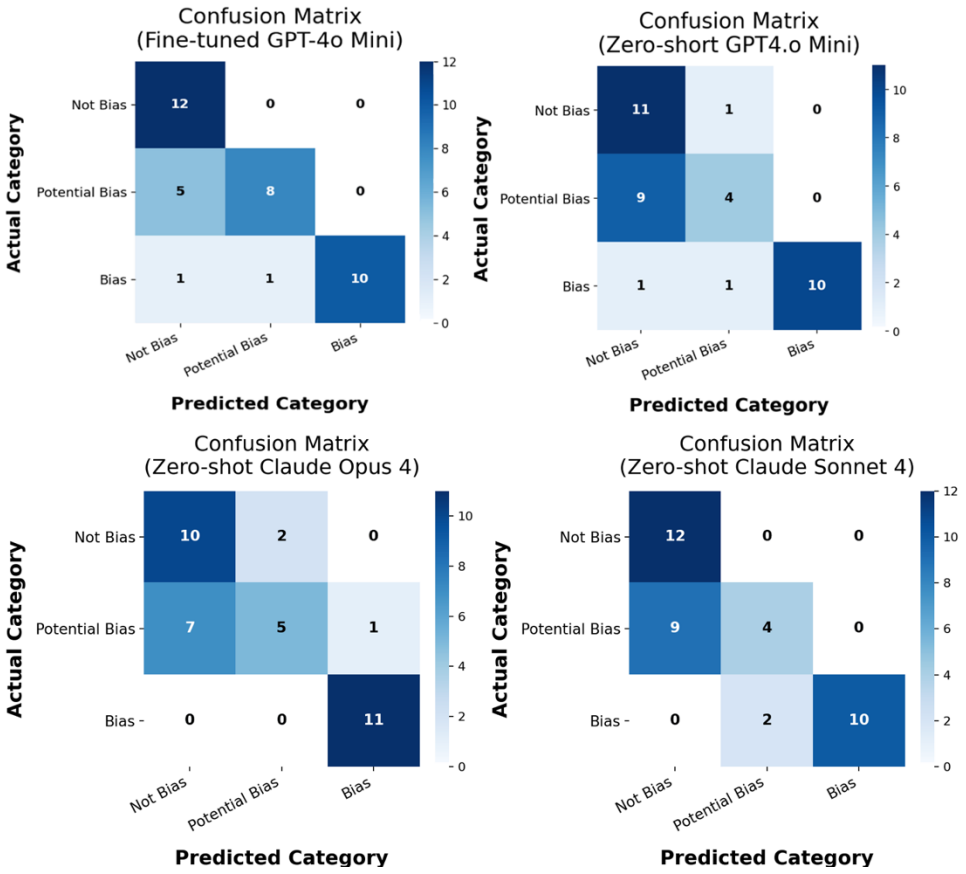

Figure 6. 1: Confusion matrices for bias class predictions for the fine-tuned GPT-4o mini and zero-shot LLM baselines (GPT-4o mini, Claude 4 Sonnet and Claude 4 Opus)

Figure 6.1 highlights clear differences across the four models. Both Claude models perform well on 'Bias' and 'Not Bias' but struggle with 'Potential Bias', often defaulting to 'Not Bias', with Sonnet showing the most conservative behaviour. The zero-shot GPT-4o mini is more balanced but still shows confusion in the 'Potential Bias' category. In contrast, the fine-tuned GPT-4o mini achieves the most consistent performance, maintaining strong results across all classes and showing the greatest improvement in handling 'Potential Bias', despite some remaining bias towards 'Not Bias'.

Focusing on the fine-tuned GPT-4o mini model, the confusion matrix reveals a clear and consistent asymmetry: errors are far more likely to occur when 'Bias' or 'Potential Bias' content is predicted as 'Not Bias' than the reverse. In contrast, cases where 'Not Bias' content is incorrectly classified as 'Bias' are not evident in this sample, indicating that the model does not tend to over-predict bias. This suggests that the model is conservative in assigning bias, favouring false negatives over false positives. While this reduces the risk of over-labelling content as biased, it also means that more subtle or context-dependent forms of bias may be overlooked.

This tendency is most evident in borderline and context-dependent cases. For example, in a historical narrative describing a legal dispute, the repeated reference to an individual as "the Indian" was annotated as 'Bias'. The annotators explicitly highlighted the

persistent emphasis on ethnicity, the use of evaluative language such as "wicked", and the implication that the individual was "not being able to pay for it" as reinforcing negative, stereotypical associations. From this perspective, the bias is not just in the label itself, but in how ethnicity is foregrounded and linked to moral judgement and financial irresponsibility.

The model, however, classified the same text as 'Not Bias' with full confidence, interpreting it as a factual recounting of events. This divergence is not simply an error but reflects a deeper disagreement about what constitutes bias in historical writing. The annotators adopt a framing-sensitive approach, treating repeated ethnic labelling and negative connotations as evidence of bias, even within a historical context. The model, by contrast, appears to prioritise narrative intent and surface factuality, discounting the implicit framing effects.

Unlike more ambiguous cases, this example is less about annotation inconsistency and more about a clear difference in evaluative criteria. The annotators are aligned and confident in identifying bias, suggesting a stable ground truth within that interpretive framework. The model's failure here therefore points to a specific limitation; it underweights how repeated descriptors and connotative language can encode bias, particularly in historical texts where such patterns are more subtle and normalised. The fact that both sides express high confidence further reinforces that confidence is not a reliable indicator of correctness in such cases, but rather of internal consistency within differing interpretive approaches.

A similar pattern emerges in several 'Potential Bias' cases that are ultimately classified as 'Not Bias', particularly where the content is educational or analytical. For example, in a classroom prompt analysing Brown's slave narrative, students are asked to consider elements such as the auctioneer's claim that a slave "has got religion" and the portrayal of slaveholders' religious behaviour. This was labelled as 'Potential Bias' by some annotators, who argued that phrases like "has got religion" and references to slaveholders' "enthusiasm for prayer" could reflect historically loaded or biased representations. However, another annotator explicitly judged the same text as neutral, emphasising that it simply encourages critical engagement with the material and does not promote a particular viewpoint.

The model classified this example as 'Not Bias', interpreting it as an instructional prompt designed to develop analytical thinking rather than to express bias. In this case, the disagreement is not primarily between the model and annotators, but among the annotators themselves. One interpretation focuses on the presence of historically charged language, while another focuses on the pedagogical function of the text. This makes the classification inherently unstable: the same phrases can be seen either as reproducing bias or as objects of critical analysis.

Unlike clearer bias cases, this example highlights the core difficulty of the 'Potential Bias' category namely, that it often depends on whether emphasis is placed on content or intent.

The model consistently defaults to interpreting such material as neutral when it is framed as educational, whereas some annotators remain sensitive to the possibility that even analytical use of language may carry implicit bias. As a result, this type of misclassification is better understood as a reflection of genuine subjectivity rather than a straightforward model failure. Notably, the model's lower confidence in this case (66.67%), compared to the previous high-confidence disagreement, suggests that confidence may serve as a useful indicator of ambiguity, signalling cases where the model is less certain and human review may be more appropriate.

Another example appears in a discussion of Malcolm X, where annotators consistently identified 'Potential Bias', pointing to phrases such as "complicated by his departure", "controversial exit" and "notoriety" as introducing subtle negative framing. The annotators argued that this language shapes the reader's perception by implying ambiguity, conflict, or inconsistency in his legacy, potentially downplaying the clarity and evolution of his later views. The model, however, classified this as 'Not Bias', interpreting the text as a neutral educational summary.

In this case, the disagreement is less about subjectivity and more about a specific limitation in the model's sensitivity to connotation and tone. While the content is informational, the framing carries evaluative weight that the model fails to capture. Unlike more ambiguous examples, all annotators converge on the presence of bias, suggesting a more stable ground truth. This highlights an important distinction: not all disagreements reflect inherent ambiguity some point to systematic blind spots, in this case the model's tendency to overlook how word choice subtly shapes interpretation.

Notably, the model's confidence here is relatively low (66.67%), which contrasts with earlier high-confidence misclassifications and suggests that it does register some uncertainty in these framing-heavy cases. This reinforces the idea that lower confidence particularly in 'Not Bias' predictions can serve as a useful indicator for cases where implicit bias may be present but not fully recognised.

Taken together, these examples show that misclassifications are not random but follow a consistent pattern: the model tends to under-detect bias where it is implicit, context-dependent, or conveyed through framing rather than explicit language. In practice, this means that both 'Bias' and 'Potential Bias' cases are sometimes resolved as 'Not Bias' reinforcing a decision boundary that favours surface neutrality unless bias is clearly stated.

The examples also demonstrate that not all errors are of the same nature. Some reflect a divergence in evaluative criteria between annotators and the model particularly around how historical language and depersonalisation should be interpreted. Others point more clearly to model limitations, specifically its reduced sensitivity to connotation and subtle linguistic framing. In contrast, some cases reveal genuine disagreement even among annotators, highlighting that certain instances are inherently subjective rather than incorrectly labelled.

This distinction is important when considering annotation quality. In some cases, particularly those involving educational content, annotations may appear overly sensitive, treating the discussion of bias as evidence of bias itself. In others, the annotations capture nuances such as tone, emphasis and implied perspective that the model systematically overlooks. As a result, disagreement should not be interpreted uniformly as error, but rather as a mix of subjectivity, annotation judgement and model blind spots.

Confidence scores further reinforce this interpretation. At a general level, they do not reliably distinguish between correct and incorrect predictions, as both high and low confidence outputs appear across cases. High-confidence disagreements, such as the historical narrative example, demonstrate that confidence reflects internal consistency rather than correctness.

However, when examined at the case level, confidence becomes more informative. Lower confidence in 'Not Bias' predictions such as in the Malcolm X and slave narrative examples often coincides with instances where the model is uncertain but defaults to neutrality. In contrast to earlier high-confidence errors, this suggests that confidence can act as a practical indicator of ambiguity, helping to identify cases where implicit bias may have been partially detected but not fully acted upon.

Overall, the confusion matrix is most useful not as a measure of aggregate performance, but as a diagnostic tool for understanding failure modes. The dominant issue is systematic under-recognition of bias in ambiguous contexts, shaped both by the model's reliance on explicit signals and by the inherent subjectivity of the 'Potential Bias' class.

6.6.2 Analysis of Bias Categories

6.6.2.1 GPT-4o mini.

A confusion matrix (see Figure 6.2) was generated to analyse the model's ability to classify bias into the correct categories. Each row shows the true bias category for each annotated example, whilst each column shows the predicted category assigned by the GPT-4o mini model. The value in each cell represents how often a true bias category was predicted as that particular category. Because some examples contain multiple bias categories, the totals in the matrix may exceed the number of examples in the dataset. Although an instance is considered correct if the model's predicted categories include at least one true category, the confusion matrix provides a clear visualisation of where the model most frequently confuses or overlaps bias types.

The findings revealed notable trends in the model's predictions. One of the most significant observations was that 'Linguistic' emerged as the most frequently predicted category. However, it also exhibited the highest number of false positives, as it was often incorrectly assigned to instances of 'Stereotype' and 'Representation'. This suggests that the model tends to overgeneralise linguistic elements of bias, even in cases where the bias is more structural or representational. Additionally, 'Stereotype' and 'Representation' were

frequently confused with each other, with many texts that should have been classified under 'Representation' being misidentified as either 'Linguistic' or 'Stereotype'. This points to a potential overlap in the linguistic features used to distinguish these categories, making it difficult for the model to differentiate between explicit stereotyping and subtler forms of representational bias.

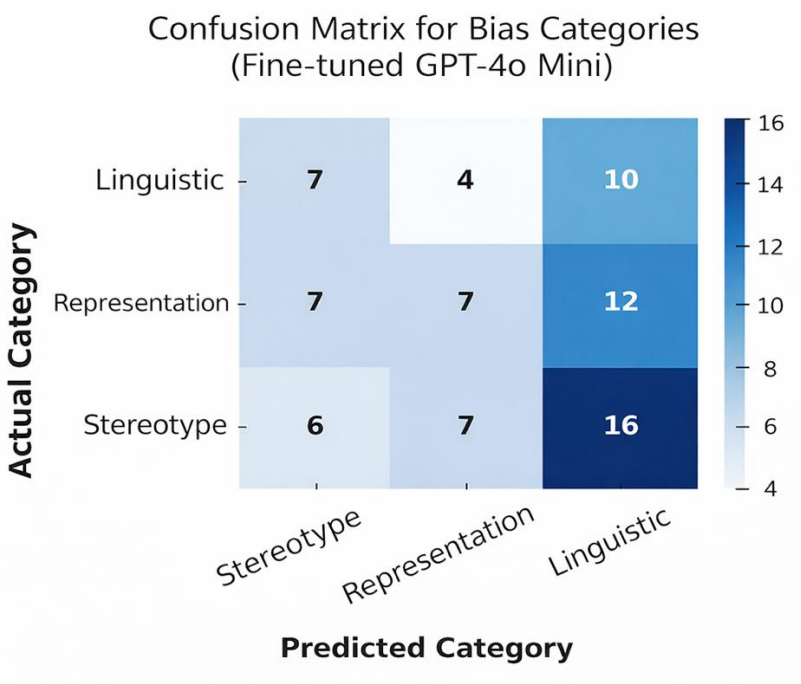

Figure 6.2: Fine-tuned GPT-4o mini confusion matrix for bias category predictions

Another notable issue was the underprediction of multi-category biases. In cases where the ground truth assigned multiple bias categories, the model often predicted only one. This suggests that the model struggles with multi-label classification, possibly due to an inability to capture the nuanced ways in which different types of bias can coexist within a single text segment. For instance, in one misclassified entry, the ground truth labelled a passage as containing both 'Representation' and 'Linguistic' bias, but the model only predicted 'Representation'. This pattern indicates that the model is more adept at identifying one dominant bias type rather than recognising instances where multiple biases intersect. The underlying issue appears to stem from overlapping linguistic markers, which may be causing the model to gravitate toward a single, more prominent category instead of capturing the full scope of bias present in the text.

Several cases exemplify this challenge, where the ground truth assigned multiple bias categories, but the model either predicted only one or an incorrect combination. This is particularly concerning in cases where the model failed to recognise representation bias, despite identifying stereotype. For example, in one instance, the text segment stated:

> "The most stable and successful governments throughout history have been led by strong, decisive male leaders, while women have traditionally focused on family and caregiving".

The ground truth classification assigned both 'Stereotype' and 'Representation' as the bias categories, as the passage not only reinforces gender stereotypes but also implicitly frames leadership as an inherently male domain, thereby marginalising the role of women in governance. The model prediction, however, only identified 'Stereotype', missing the broader representational aspect of the bias. While the model correctly flagged the stereotype that men are more suited to leadership, it failed to capture the systemic implication that women are inherently unsuited for such roles. This oversight indicates that the model is better at detecting overt stereotyping but less proficient at identifying structural and representational biases, which are often more implicit and embedded in the framing of statements.

Another recurring misclassification involved the overgeneralisation of 'Linguistic' as a category. The model frequently misclassified texts that should have been categorised under 'Representation' or 'Stereotype' as belonging to 'Linguistic', likely due to framing choices or the use of emotionally charged language. One such example is the statement:

> "The rapid development of Western economies can be attributed to their superior innovation and work ethic, which contrasts with the slower economic progress seen in other regions".

The ground truth classification assigned both 'Representation' and 'Stereotype', as the passage suggests that Western success is due to inherent superiority rather than historical, economic, or political factors, reinforcing a bias in how economic progress is framed. However, the model classified it as 'Stereotype' and 'Linguistic', misidentifying the representational bias as a matter of tonal framing rather than a systemic portrayal of Western superiority. This misclassification suggests that the model struggles to distinguish between biases that stem from portrayals of group superiority (Representation Bias) and biases that rely on explicit stereotypes (Stereotype Bias). The phrase "superior innovation and work ethic" may have also triggered the 'Linguistic' classification due to its evaluative tone, reinforcing the model's tendency to rely on word-level cues rather than deeper contextual meaning.

Another common challenge the model faced was the ambiguity in distinguishing between 'Representation' and 'Stereotype'. Biases embedded within historical or cultural narratives were often misclassified between these two categories, suggesting that the model lacks the ability to differentiate between broad societal portrayals and explicit stereotyping. This pattern can be observed in the following example:

> "Crime rates tend to be higher in urban areas due to cultural differences and lack of traditional family structures".

The ground truth classification assigned both 'Stereotype' and 'Linguistic', as the passage implies a stereotypical association between crime and cultural differences, reinforcing a negative framing of certain social groups. Additionally, the phraseology introduces an implicit negative tone, influencing how the information is perceived. However, the model predicted 'Stereotype' and 'Representation', incorrectly identifying 'Representation' instead of 'Linguistic'. This misclassification reveals two important trends.

Firstly, the model appears to be better at recognising explicit stereotypes it correctly identified that the passage reinforces a generalisation about crime and cultural differences. However, it struggled to detect biases hidden in linguistic framing and tonal shifts, leading to the misclassification of 'Linguistic' as 'Representation'. This suggests that the model relies more on the subject matter and explicit content rather than recognising how the tone or word choice influences bias perception.

Secondly, the misclassification of 'Linguistic' as 'Representation' suggests that the model may conflate portrayals of social groups with language-based bias markers. While 'Representation' typically refers to how groups are depicted in media, discourse, or historical narratives, the model's classification indicates that it often assigns this category to statements that negatively frame certain communities, even when the bias is more about tone and language than group depiction. This implies that the model requires further refinement in distinguishing between content-driven bias (Representation) and linguistic framing bias (Linguistic).

In conclusion, the model demonstrates strong performance in detecting bias and effectively categorises many instances with high accuracy. It performs particularly well in identifying explicit biases and consistently recognises key bias patterns across different texts. While some challenges remain in distinguishing between 'Representation', 'Stereotyping', and 'Linguistic', the model still provides valuable and insightful classifications. Its ability to capture bias in a wide range of contexts highlights its effectiveness as a tool for bias detection, offering a solid foundation for further refinement and application in higher educational texts within the humanities.

6.6.2.2 Zero-Shot GPT-4o mini.

The model shows reasonable capability in detecting explicit 'Stereotype' when bias indicators are overtly present in the text, such as direct gender role assignments or ethnic generalisations. However, it struggles considerably with 'Representation' bias, often failing to recognise when groups are marginalised through omission or limited portrayal rather than explicit negative characterisation. The most striking limitation of the zero-shot model is its difficulty with 'Linguistic' detection. Unlike the fine-tuned version, which tends to over-predict this category, the zero-shot model rarely identifies subtle linguistic bias markers. This suggests that recognising implicit tonal bias requires the nuanced pattern recognition that emerges through fine-tuning on domain-specific examples (see Figure 6.3).

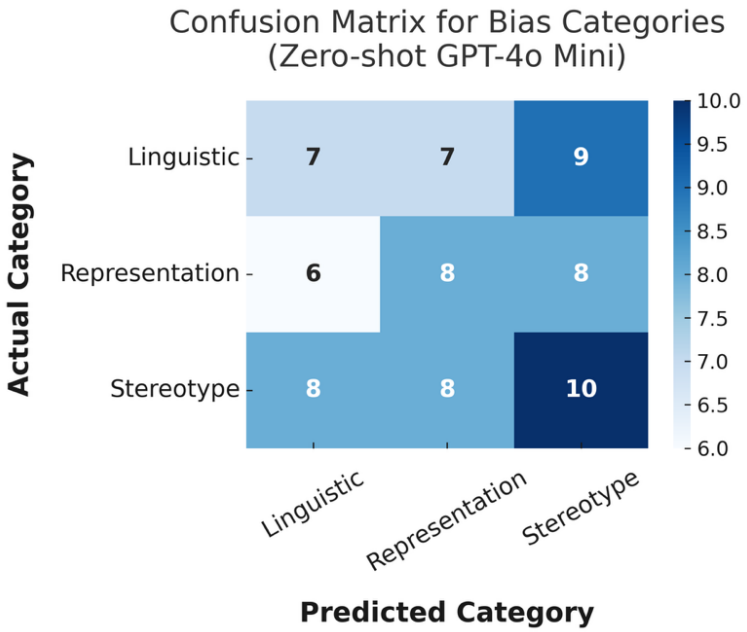

Figure 6.3: Zero-shot GPT-4o mini confusion matrix for bias category predictions

6.6.2.3 Zero-Shot Claude Models.

As with the fine-tuned model, Figure 6.4 presents the multi-label confusion matrix for zero-shot predictions. The main trends are consistent: 'Linguistic' is the most frequently predicted category and is also most prone to false positives, often being incorrectly assigned to examples that should be 'Stereotype' or 'Representation'. This pattern suggests that the zero-shot Claude models, like the fine-tuned GPT-4o mini, tend to overgeneralise linguistic markers of bias, sometimes conflating word choice or tone with deeper structural or representational bias. The confusion between 'Stereotype' and 'Representation' remains pronounced. For instance, the models often mislabel examples that involve marginalisation of groups 'Representation' as simple 'Stereotype', indicating difficulty in distinguishing between overt stereotypes and more implicit forms of exclusion or omission. Furthermore, multi-category biases are frequently under predicted: when the ground truth includes both 'Representation' and 'Linguistic', the zero-shot models typically predict just one category. Overall, while the zero-shot models are effective at detecting explicit bias, they share similar limitations to the fine-tuned model in handling subtle or intersecting bias categories. This demonstrates the need for further prompt engineering or model refinement to improve nuanced, multi-label bias detection in educational texts.

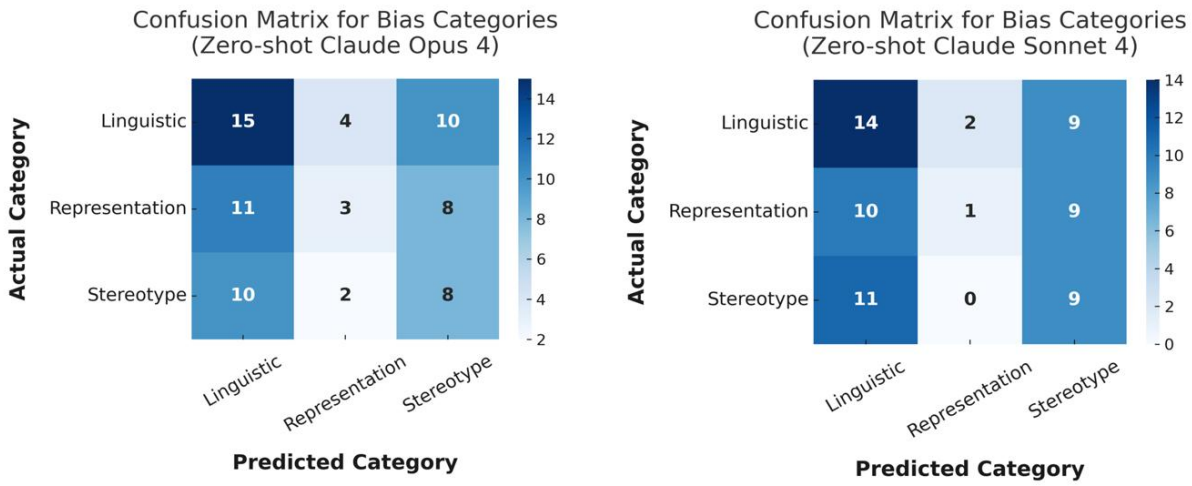

Figure 6.4: Zero-Shot LLM baselines Claude-4's Sonnet and Opus confusion matrix for bias category predictions

6.7 Subjectivity in Annotation and Fine-tuned GPT-4o mini Validation

The annotation of bias in the OER-H dataset is highly subjective, as reflected by low agreement among annotators discussed in Chapter 3 section 3.6.2. To provide a clearer evaluation, a subset of the test set where all three annotators agreed on a single class label is considered. These cases represent the highest level of certainty, where bias is least ambiguous. Out of 37 test cases, only 8 achieved full consensus, highlighting the difficulty of establishing a clear ground truth.

On this high-confidence subset, the fine-tuned GPT-4.0-mini model agreed with annotators in 2 cases, corresponding to 25% agreement. While this appears low, it is important to note that the majority of the dataset is inherently ambiguous, limiting the reliability of standard accuracy metrics. The model's correct predictions in these clear cases suggest it can capture some patterns of bias, but overall evaluation is constrained by the subjectivity of the annotations.

6.8 Cross-Domain Validation

To assess model generalisability, the fine-tuned GPT-4o mini and zero-shot GPT-4o mini were each applied to UoL-DSPDS-LT dataset. This unannotated dataset was chosen to examine whether bias detection methods developed on OER-H could identify bias within technical disciplines traditionally regarded as more objective.

6.8.1 Zero-Shot and Fine-Tuned GPT-4o mini Performance on STEM Content

Both models exhibited very low rates of bias detection in the dataset. The fine-tuned GPT-4o mini classified 221 out of 223 segments (99.1%) as Not Bias, flagged one segment (0.4%) as biased, and marked one as Unknown. The presence of an 'Unknown' label is a noteworthy behaviour, as the model had access to a Potential Bias class that would

typically be used in cases of uncertainty. This may indicate model hesitation when encountering a borderline case, reflecting a conservative bias-flagging strategy.

The zero-shot GPT-4o mini showed slightly greater sensitivity, identifying one 'Bias' instance (0.4%) and three 'Potential Bias' instances (1.3%), with the remaining 219 segments (98.2%) classified as neutral.

The single bias instance identified by both models involved a lecture segment discussing data visualisation scales, where the lecturer asked whether "people from India are quite short" based on a misleading graph scale. Both models correctly recognised this as representational bias namely, an ethnic generalisation even though it was embedded in critique of the data presentation.

The text segment:

> "The information says the people from India are quite short, and is that true actually? Because the scale is starting from five feet and it's only this bit there, so definitely they are not this short compared to this one from Latvia. We need to be careful of what we want to say from our graphs, and it has to be accurate. It's giving us the idea that they are really short in India, but they are not actually, it's just the scale that we are using that's making this look like that."

The zero-shot model flagged three additional segments as 'Potential Bias' that the fine-tuned model classified as neutral:

- **Medical Categorisation:**

> "Sometimes the data has missing data… there are some cancer diseases that are specifically for men and others that are more common in women. So, if you try to combine… there will be missing fields for females who wouldn't get the prostate cancer for instance and vice versa…"

This was flagged as reinforcing binary gender assumptions in medical data, potentially excluding non-binary or intersex identities.

- **Gender Binary Assumptions (Imputation):**

> "One student in the last run suggested why don't we get the mean of the females and the mean of the males and impute them. If the missing age record is for a female, then we impute it with

> the mean age of females and if it is for male, then we impute it with the mean of the male."

This was identified as perpetuating binary gender categorisation in data imputation strategies.

- **Educational Value Discourse:**

> "This is my favourite actually. It's saying that there is no value in doing college education. It's correct, it's starting from the right point… but it's just showing one side of the story… You are trying to say that it's not worth it doing higher education…"

The zero-shot model flagged this for one-sided framing of educational value, which may implicitly dismiss alternative perspectives.

The very low bias detection rates (between 0.4 % and 1.8 % across models) suggest that STEM academic content may indeed contain substantially less explicit bias than humanities materials, which is consistent with traditional assumptions of objectivity within STEM disciplines. STEM subjects often rely on quantitative methods, standardised technical terminology, and empirical or procedural objectivity, which may reduce (but not eliminate) opportunities for explicit social or representational bias to appear in discourse (Charlesworth and Banaji, 2019).

Nevertheless, the detection of at least one clear bias instance and the additional borderline cases identified by the zero-shot model shows that bias can still emerge in technical teaching materials, particularly through representational and categorical framing. The fine-tuned model's use of an Unknown label, rather than 'Potential Bias', might indicate reluctance to over-predict bias outside its training domain, highlighting a limitation in cross-domain sensitivity. In contrast, the zero-shot model's willingness to use 'Potential Bias' suggests better adaptability to ambiguous cases.

6.9 Limitations

This chapter acknowledges several important limitations that should be considered when interpreting the findings.

6.9.1 LLM-Only Approaches and Nuanced Bias Detection

The most important limitation concerns the marked constraints of LLM-only approaches in handling nuanced and context-dependent bias. Despite recent advances in automatic bias detection using LLMs, this study demonstrates that such approaches still face some challenges. The fine-tuned GPT-4o mini model consistently struggled to distinguish between bias categories, frequently confusing 'Stereotype', 'Representation' and 'Linguistic' bias. The confusion matrices revealed that the model overgeneralised 'Linguistic' bias,

misclassifying instances that should have been identified as other categories, suggesting overlapping linguistic features that the model cannot adequately differentiate.

More critically, the model failed to identify instances involving multiple or overlapping bias categories. When ground truth assigned multiple bias categories to a single segment, the model typically predicted only one, missing the intersectional nature of bias. This limitation is particularly problematic since educational bias frequently manifests through multiple, simultaneous mechanisms. For instance, a passage may employ stereotypical language whilst also marginalising groups through limited representation and framing issues through biased linguistic choices, yet the model identifies only the most dominant pattern.

The 'Potential Bias' class presented the most pronounced challenges. Even the fine-tuned model achieved only 0.62 recall for this class, missing nearly two-fifths of instances. This category, designed to capture ambiguity and subjectivity where bias depends heavily on context and interpretation, consistently achieved lower recall and F1 scores across all models. The frequent misclassification of 'Potential Bias' as 'Not Bias' is particularly concerning for educational applications, where overlooking subtle bias could allow problematic content to remain undetected.
The complexity of educational texts in humanities further compounds these difficulties. These domains are characterised by contested narratives, multiple viewpoints, and frequent interpretive disagreement. Statements on topics such as gender representation, colonial histories, and national identity may be perceived differently depending on context, reader background, and cultural perspective. This inherent interpretive variability presents fundamental challenges for both human annotators and automated systems, highlighting the need for approaches that can support more context-sensitive, interpretable, and robust bias detection rather than relying solely on LLM classification.

6.9.2 Dataset Size Limitation

The OER-H dataset's small scale presents additional challenges. With only 183 entries total (108 original, 75 synthetic) and a test set of just 37 examples, performance metrics must be interpreted cautiously. The limited size raises concerns about generalisability, and the dataset's focus on humanities disciplines constrains applicability to specific academic contexts.

The use of synthetic data augmentation for training, whilst methodologically sound in preventing data leakage by keeping augmented examples separate from the test set, means the model was exposed to GPT-4o-generated patterns during learning. These augmented examples, though designed to align with original data patterns, may not fully capture authentic educational materials' nuanced complexity and could have introduced artificial patterns that influenced the model's learning behaviour.

6.10 Conclusion

This chapter demonstrated that fine-tuned GPT-4o mini achieves 81.1% accuracy in detecting bias within OER-H, substantially outperforming zero-shot baselines. The model performs well on explicit bias detection but struggles with nuanced 'Potential Bias' categories and over-predicts 'Linguistic' bias markers.

Cross-domain validation revealed minimal bias detection in technical disciplines (0.4-1.8%), with both models identifying only one common biased segment involving ethnic generalisation. This suggests either genuine objectivity in technical content or limited cross-domain generalisability.

Key contributions include a comprehensive annotation framework achieving 81.43% inter-annotator agreement and demonstration that modest data augmentation enhances performance on small, specialised datasets. Areas for development include expanding dataset coverage, improving multi-label classification, and refining bias category distinctions.

The findings suggest AI-based bias detection offers promising support for developing more inclusive educational content when implemented with appropriate human oversight. The following chapter continues this line of investigation, utilising the same datasets and models while introducing a hybrid RAG approach to enhance analytical depth and contextual understanding.

Chapter 7

7 Hybrid RAG and Fine-Tuned LLM for Bias Detection in Online Higher Education Learning Resources in Humanities

7.1 Introduction

Recent advances in LLMs have significantly improved the automatic detection of bias in educational texts (Fan, Z. et al., 2024; Lin et al., 2024; Weissburg et al., 2024; Ahmed, I. et al., 2025; Albuquerque et al., 2025). Nevertheless, as demonstrated in the preceding chapter, LLM-only approaches still display marked limitations, particularly in handling nuanced and context-dependent cases of bias. The fine-tuned GPT-4o mini model on OER-H dataset, for instance, frequently confuses between bias categories and often fails to identify instances involving multiple or overlapping bias categories (see Chapter 6 section 6.6.2). These limitations are especially pronounced in the case of the 'Potential Bias' bias class, which is intended to capture ambiguity and subjectivity, but which consistently achieves lower recall and F1 scores across models.

The complexity of educational texts in humanities further complicates reliable bias detection. These domains are characterised by contested narratives, multiple viewpoints, and frequent interpretive disagreement (McCullagh, 2000; Bhat et al., 2023; Kropman et al., 2023; Rist, 2023). Statements on topics such as gender representation, colonial histories, and national identity may be perceived differently depending on context and background knowledge, leading to inherent challenges for both human annotators and automated systems. As such, there is a clear need for approaches that can support more context-sensitive, interpretable, and robust bias detection.

One promising direction, supported by recent work in NLP, is the use of RAG (Lewis, P. et al., 2020). This technique supplements the predictive power of LLMs with annotated examples from an external corpus, allowing each decision to be informed by relevant, domain-specific precedents. Retrieval-based methods have been shown to improve both performance and interpretability, particularly on complex or ambiguous tasks where a single model's pre-trained knowledge may not suffice (Zhao, S. et al., 2024; Li et al., 2025; Ren et al., 2025).

This chapter presents a hybrid bias detection system that integrates a fine-tuned GPT-4o mini with a RAG component. The hybrid architecture addresses the shortcomings of LLM-only systems by exposing the model to a diverse range of annotated precedents at inference time, promoting more balanced and transparent predictions.

7.1.1 Objectives of the Chapter

The objective of this chapter is to design and implement a hybrid bias detection system that integrates a fine-tuned GPT-4o mini model with a RAG component, specifically tailored for detecting bias in higher educational Humanities materials within the OER-H dataset.

7.1.2 Contributions of The Chapter

This chapter aims to address the third research contribution outlined in Chapter 1, Section 1.3.3.4, designing a hybrid framework that integrates a fine-tuned LLM model with RAG for bias detection.

7.2 RAG Corpus Details

The retrieval corpus consists of 220 annotated educational text segments. Importantly, these specific data entries were excluded from the fine-tuning and testing of the GPT-4o mini model introduced in the previous chapter to maintain a strict separation between the data used for fine-tuning and the retrieval corpus. The corpus is composed of:

- 182 examples originate from the same source dataset (i.e., OER Commons[23]) introduced in Chapter 3. However, 150 entries were annotated using GPT-4o and kept distinct from OER-H. The remaining 32 entries were drawn from the subset re-annotated by human annotators on the Prolific platform. Crucially, these 32 human-annotated entries were excluded from OER-H (More information about the data and the annotation framework can be found in Chapter 3 Sections 3.3.3 and 3.6.).

- 38 newly augmented domain-specific entries generated using GPT-4o through the data augmentation methodology detailed in Chapter 6 Section 6.3. This process involved prompting GPT-4o to create new, realistic text segments that align with both the topical scope and annotation framework of the primary dataset. The augmentation strategy was designed to enhance the retrieval corpus by expanding its topical coverage and contextual diversity, thereby improving the system's ability to retrieve relevant examples across a broader range of educational content scenarios. To prevent data leakage, these augmented entries were derived exclusively from the 213 RAG corpus entries in the previous point and were verified to have no overlap or similarity with the data used for fine-tuning or testing GPT-4o mini. Each entry is labelled with bias class, bias category (when relevant), keywords, and a justification, following the same annotation schema described in Chapter 3 section 3.6.

[23] https://oercommons.org/

7.3 RAG Baseline

Prior to implementing the hybrid approach, an experimental baseline was established using retrieval alone, without involving the generative model. Under this experimental baseline, classification is performed using a nearest-neighbour majority vote procedure (Cover and Hart, 1967). The majority vote classification procedure for the RAG baseline can be described as follows:

Given a test input x, the system retrieves the k most similar annotated examples from the corpus, denoted as $[\mathrm{y}_1, \mathrm{y}_2, \mathrm{y}_3, \ldots, \mathrm{y}_\mathrm{k}]$, each associated with a class label $\mathrm{c}(\mathrm{y}_\mathrm{i})$. The predicted class $\hat{c}(x)$ for the input x is then determined by selecting the most frequently occurring class label among these retrieved neighbours:

$$\hat{c}(x) = arg\max_{c \in C} \sum_{i=1}^{k} I(c(y_i) = c)$$

where:

- c is the set of possible bias class labels (Bias, Potential Bias, Not Bias).
- $I(.)$ is the indicator function, defined as:

$$I(z) = 1 \text{ if } z \text{ is true; otherwise } 0$$

Thus, the input is classified into the category most commonly represented among the nearest neighbours.

Under this retrieval-only framework, each input sentence is first encoded as a dense vector representation using the pre-trained sentence transformer model 'all-mpnet-base-v2' (Reimers and Gurevych, 2019; Song et al., 2020). Cosine similarity between the input embedding and the precomputed embeddings of all retrieval entries is computed using the formulation previously presented in Chapter 5.

The top-k most similar annotated entries are identified based on their similarity scores. The value of k was determined through experimentation with multiple settings (e.g., k = 3, 5, and higher), with k = 3 yielding the most reliable results. Increasing k beyond this point introduced less relevant entries and reduced retrieval precision. Accordingly, k = 3 was consistently adopted. The predicted class for each input text is determined by selecting the most frequent bias class among these k retrieved neighbours. This simple, interpretable approach establishes a clear baseline for assessing how effectively retrieval alone can handle the classification task without the involvement of any generative components.

7.4 Hybrid RAG and Fine-Tuned LLM Architecture

The hybrid bias detection system integrates the fine-tuned GPT-4o-mini with a RAG module, forming a unified pipeline for classifying bias in educational texts. This architecture is designed to address the limitations observed in LLM-only approaches, specifically by enabling each prediction to be informed by relevant, domain-specific precedents. At a high level, the system comprises two main components:

1. **Fine-Tuned LLM Model**: The GPT-4o mini model introduced in the previous chapter, which is fine-tuned on OER-H. At inference time, the model receives user input and generates structured outputs, including bias class, bias category, relevant bias keywords, a justification of the bias decision and a confidence score. Model outputs are parsed using a repair function to ensure consistency and robustness.

2. **Retrieval Module**: The RAG module serves as the retrieval component of the system, responsible for providing the fine-tuned GPT-4o mini model with relevant, contextually grounded examples at inference time. When a user provides an input excerpt, it is first encoded into a vector representation using a sentence transformer, after which the retrieval module selects the k–3 most similar annotated examples from the RAG corpus based on cosine similarity. The top-k most similar annotated entries are identified based on their similarity scores. The value of k was determined through experimentation with multiple settings (e.g., k = 3, 5, and higher), with k = 3 yielding the most reliable results. Increasing k beyond this point introduced less relevant entries and reduced retrieval precision. Accordingly, k = 3 was consistently adopted for both the evaluation process and the end-user RAG system. These retrieved examples are incorporated into the model prompt in a structured format that mirrors the annotation schema, supplying the model with explicit precedents to guide its prediction. This process improves the consistency, interpretability, and domain coverage of the model's bias classifications while allowing the retrieval corpus to be updated independently of the model's parameters, thereby keeping the system adaptable over time.

3. **Post-processing and Validation of Model Outputs:** To ensure the reliability and interpretability of model outputs, a post-processing and validation mechanism was implemented within the application pipeline. After the model generated a response in JSON format, the output was parsed into a structured Python dictionary and subjected to a series of preprocessing steps. These steps included textual normalisation, where both the input text and extracted keywords were cleaned by removing punctuation and converting to lowercase, and structural standardisation, ensuring consistency in fields such as classification labels and categories. A key component of this process involved filtering the model's extracted keywords. Only bias keywords or phrases that could be directly matched to the original input text were retained, thereby preventing the inclusion of unsupported or hallucinated evidence. Additionally, when the model classified a text as 'Not Bias',

any associated keywords were removed to maintain internal consistency. Following preprocessing, a validation step was applied to enforce a minimum evidentiary standard. Specifically, any output labelled as 'Bias' or 'Potential Bias' was required to include at least one supporting keyword or phrase. Outputs that failed to meet this requirement were treated as invalid. In such cases, the system automatically reissued the model request, up to a maximum of three attempts, with an augmented instruction emphasising the need for explicit evidence. This post-processing and validation mechanism functioned as a quality control layer, improving the consistency, transparency, and trustworthiness of the model's classifications before they were stored and used within the application.

A schematic of this architecture is presented in Figure 7.1, illustrating the complete data flow from user input through retrieval, prompt augmentation and final output validation.

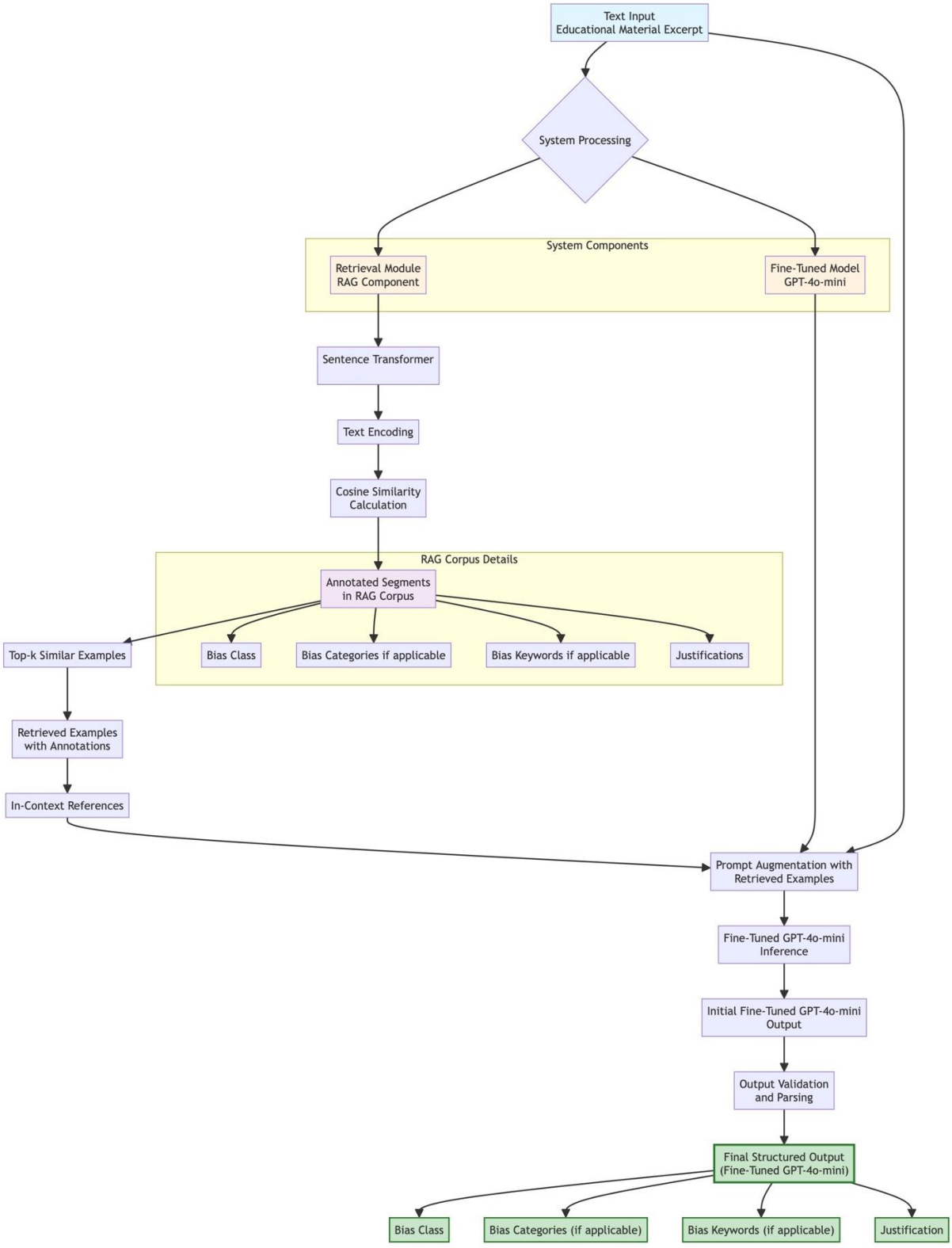

Figure 7.1: Architectural overview of the hybrid bias detection system integrating RAG and fine-tuned GPT-4o mini

Based on Figure 7.1 the hybrid architecture combines a RAG component with a fine-tuned GPT-4o mini model to detect bias in higher education content. When a user provides an input excerpt, it is first encoded into a dense vector representation using a sentence transformer. The retrieval module then performs a cosine-similarity search against the RAG corpus and selects the k (i.e., k=3) most relevant annotated examples. These retrieved examples containing bias classes, categories, keywords, and justifications are inserted into the model prompt in a structured format, giving the fine-tuned model explicit, contextually grounded precedents to consider.

Crucially, while the retrieved examples influence the model's reasoning by providing additional context, the final prediction is always determined solely by the fine-tuned GPT-4o mini model**.** After inference, a validation and parsing step ensures that the output conforms to the required schema, correcting any inconsistencies when necessary.

To illustrate the operation of the hybrid system, consider the following representative input from the educational materials corpus:

> "Throughout history, most significant political achievements have been made by men, while women have largely remained in domestic roles."

Upon receiving this input, the retrieval module encodes the text and identifies the top k most similar annotated examples (i.e., k=3) from the RAG corpus. These retrieved examples, along with their associated bias class, category, keywords, and justifications, are appended to the model prompt as in-context references. Examples from the RAG corpus:

- **Example 1:**

> **Text Segment:** "Historically, men have held most leadership roles, while women focused on homemaking"
> **Bias Class:** Bias
> **Bias Category:** Stereotype
> **Bias Keywords:** men, leadership, women, homemaking
> **Justification:** The text assigns traditional roles by gender, reinforcing stereotypes.

- **Example 2:**

> **Text Segment:** "The most influential philosophers have been male."
> **Bias Class:** Bias

> **Bias Category:** Representation
> **Bias Keywords:** philosophers, male, influential
> **Justification:** The example highlights the underrepresentation of women in philosophy, focusing only on male figures.

- **Example 3:**

> **Text Segment:** "Men built the modern world, according to the text."
> **Bias Class:** Bias
> **Bias Category:** Stereotype
> **Bias Keywords:** men, modern world, built
> **Justification:** The statement reinforces the notion that societal progress is primarily attributable to men.

These retrieved examples, including their full annotations, are then incorporated as in-context references within the prompt presented to the fine-tuned GPT-4o mini model. The prompt is constructed as follows:

> **[Instruction]:** You are an expert bias detector. Here is the educational text to analyse:
> **[Input]:** Throughout history, most significant political achievements have been made by men, while women have largely remained in domestic roles.
> **[Instruction]:** Here are similar, annotated examples:
> **[Retrieved Example 1]:** ...
> **[Retrieved Example 2]:** ...
> **[Retrieved Example 3]:** ...
> **[Instruction]:** Using the above, classify the input as Bias, Not Bias, or Potential Bias...etc. Please output your answer in the following format:
> - **Bias Class:** ...
> - **Bias Category:** ...
> - **Bias Keywords:** ...
> - **Justification:** ...

The fine-tuned GPT-4o mini model uses the input and the retrieved examples to generate:

> **Bias Class:** Bias
>
> **Bias Category:** Stereotype
>
> **Bias Keywords:** men, political achievements, women, domestic roles
>
> **Justification:** The text attributes political achievements exclusively to men and portrays women in domestic roles, reflecting gender stereotypes.

7.4.1 Decision Logic and Fallback Mechanisms

A central consideration in hybrid bias detection systems is how to determine whether to trust the fine-tuned GPT-4o mini's own prediction or to defer to annotations from closely related, previously labelled examples provided by RAG. In this work, a similarity weighted voting framework was implemented to address this challenge.

Under this approach, the system compares each input to the set of retrieved examples from the RAG corpus, using cosine similarity to identify the most relevant precedent. If the similarity between the input and its closest retrieved example exceeds a specified threshold (e.g., 0.7), the annotation associated with this precedent is adopted as the system's prediction. It is important to note that, within the hybrid system, 'adoption' does not mean copying the annotation labels verbatim from the corpus. Rather, as illustrated in the example above, the fine-tuned GPT-4o mini model is explicitly guided by the retrieved precedent from RAG during prediction, generating its own structured output informed by the combined context. In cases where the similarity falls below the threshold, the output produced by the fine-tuned GPT-4o mini model alone is used.

This threshold-based approach operates on the principle that similarity scores above 0.7 typically indicate either paraphrased content or semantically equivalent cases, justifying direct annotation inheritance. The high threshold ensures that only genuinely analogous examples trigger precedent adoption, whilst preserving the language model's reasoning capabilities for novel, ambiguous, or contextually distinct instances. This dual-pathway mechanism maintains system robustness by leveraging exact matches when available, whilst ensuring that the fine-tuned model handles cases requiring interpretative reasoning or encountering previously unseen patterns.

7.5 Hybrid Approach Results

Table 7.1 presents a summary of the three bias detection pipelines evaluated in this study: the RAG-only baseline, the fine-tuned GPT-4o mini baseline, and the proposed hybrid approach. The performance was assessed using two metrics: relaxed accuracy and macro-F1. Relaxed accuracy as explained in the previous chapter requires that a prediction be considered correct only if both the bias class (Bias, Not Bias, Potential Bias) and at least one of the true bias categories (Stereotype, Representation, Linguistic) are predicted correctly. Macro-F1 was selected over standard accuracy because it provides a class-balanced measure of model performance, giving equal weight to each class regardless of frequency. Unlike accuracy, which may appear high even when a model performs poorly on certain classes, Macro-F1 reflects how well the model performs across all classes by averaging their individual scores:

$$Macro - F1 = \frac{1}{C}\sum_{i=1}^{C} F1_i$$

Where C is the number of classes and $F1_i$ is the score for class i. This averaging ensures that underperformance in a single class meaningfully lowers the overall score.

Macro-F1 highlights weaknesses in harder to predict classes, rather than allowing majority or easy classes to dominate the evaluation. This was particularly important for our study, as the previous experiment results showed that the fine-tuned GPT-4o mini model frequently misclassified or failed to identify instances of the 'Potential Bias' bias class; a class designed to capture ambiguous or subjective cases. Consequently, macro-F1 offers a fairer and more diagnostic evaluation metric for bias detection, revealing systematic weaknesses that plain accuracy would obscure (Opitz, 2022; Takahashi et al., 2022; Opitz, 2024).

	Embedding model (dimension)	Override Rule	Relaxed Accuracy	Macro-F1
Baseline - RAG	All-mpnet-base-v2 (768) (Song et al., 2020)	Top 3 nearest neighbours by cosine similarity; predicted label is determined by majority vote	62%	60%
Baseline - fine-tuned GPT-4o mini	—	—	81%	81%
Hybrid Approach	All-mpnet-base-v2 (768) (Song et al., 2020)	For each input, retrieve 3 most similar RAG examples; the fine-tuned GPT-4o mini predicts class using both input and RAG context	78%	78%

Table 7.1:Architectural overview and performance of baselines and hybrid models

For the RAG baseline, the predicted class was determined by majority vote among the top three most similar annotated examples, as measured by cosine similarity in the embedding space. The hybrid approach incorporated RAG, in which the k-nearest annotated examples were appended to the input prompt for the fine-tuned GPT-4o mini model; the model then produced the final prediction based on both the input and the retrieved context. While the fine-tuned GPT-4o mini achieved a slightly higher overall accuracy of 81% compared to the hybrid approach's 78%, a comprehensive evaluation requires further analysis of performance across individual bias classes and categories to determine the most appropriate approach for bias detection applications.

Moreover, the strict accuracy introduced in Chapter 6 was applied here, whereby a prediction is only considered correct if both the bias class and all associated bias categories are accurately identified. Under this criterion, the hybrid approach achieved a strict accuracy of 56.8%, which is marginally higher than the fine-tuned GPT-4o mini model alone (56.76%). While the improvement is slight, it indicates that incorporating retrieval provides a small but measurable gain in exact category prediction.

7.5.1 Hybrid Approach Bias Class Prediction

To further evaluate the hybrid system performance, a confusion matrix was constructed for the bias classes: Bias, Not Bias, and Potential Bias (see Figure 7.2).

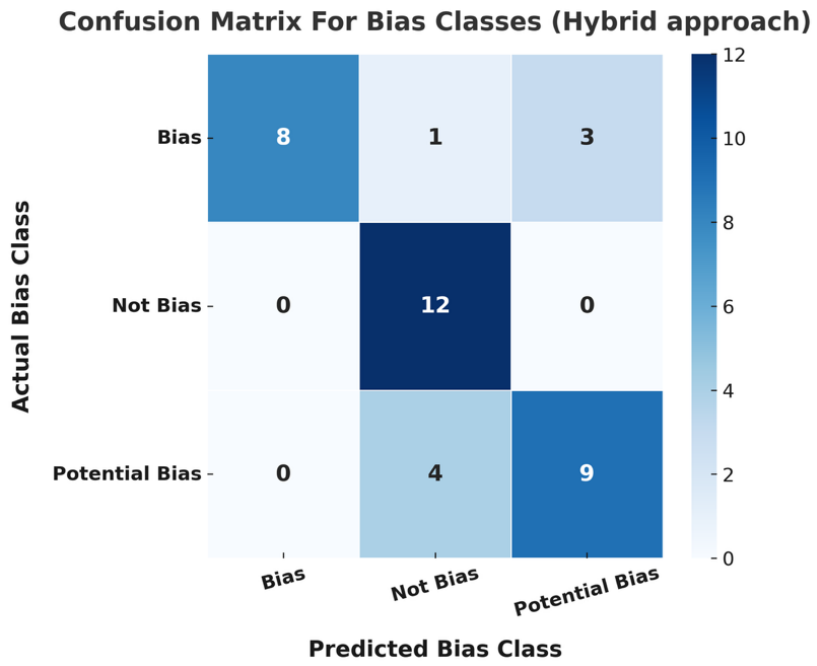

Figure 7.2: Hybrid approach confusion matrix for bias classes

This matrix indicates that the system performs strongly in identifying 'Not Bias' instances, with all twelve true negatives correctly classified and no instances of 'Not Bias' being confused for either 'Bias' or 'Potential Bias'. For the 'Bias' class, the majority of cases were correctly identified, with a small number misclassified as 'Not Bias' and several more labelled as 'Potential Bias'. This suggests that while the model is generally reliable in detecting explicit bias, there remains some tendency to conflate explicit and potential bias, reflecting the inherent subjectivity and ambiguity in borderline cases.

For 'Potential Bias', performance is more mixed: while 9 out of 13 cases were correctly identified, 4 were misclassified as 'Not Bias'. Notably, there were no cases where 'Potential Bias' was misclassified as explicit 'Bias', implying that the model is relatively conservative in assigning the explicit bias label.

Overall, the confusion matrix highlights two key trends: (1) the system is most precise in identifying text with no bias, and (2) ambiguity primarily arises between the explicit 'Bias' and 'Potential Bias' classes, as expected given the nuanced and context-dependent nature of these categories.

A comparison with the fine-tuned LLM-only approach (i.e., fine-tuned GPT-4o-mini) highlights several targeted improvements. Most notably, the hybrid system fully eliminates false positives in the 'Not Bias' class, correctly classifying all neutral instances and addressing the precision limitation observed in the fine-tuned LLM-only results (0.67 precision for this class; see Chapter 6, Table 6.3). For explicit 'Bias', performance is broadly comparable, though the hybrid model continues to occasionally reclassify clear bias as

'Potential Bias', reflecting its cautious treatment of borderline cases rather than a strict accuracy gain. Performance on 'Potential Bias' shows a moderate improvement, with the hybrid system correctly identifying a majority of these cases an enhancement over the fine-tuned LLM-only F1-score of 0.73 though several remain misclassified as 'Not Bias'.

Both approaches show a conservative pattern overall, with neither misclassifying 'Potential Bias' as explicit 'Bias', suggesting consistent reluctance to escalate from implied to overt bias class. While the hybrid approach reduces false positives and slightly improves coverage of nuanced bias, the persistent confusion between 'Bias' and 'Potential Bias' suggests that separating the two is still a challenge

7.5.2 Hybrid Approach Bias Category Analysis

Based on the confusion matrix (see Figure 7.3) analysis, the hybrid model demonstrates moderate success in distinguishing between different types of bias, though with notable variations in performance across categories. 'Representation' bias achieves the highest number of correct classifications, followed by 'Linguistics' bias, while 'Stereotype' bias shows the most challenging performance with fewer correct identifications. The hybrid approach showed a tendency to misclassify other bias types as 'Representation' bias, with 6 instances of 'Linguistics' bias and 4 instances of 'Stereotype' bias being incorrectly predicted as 'Representation'. This pattern suggests that 'Representation' bias markers may be more prominent in the model's decision-making process or that this category serves as a default classification. These misclassification patterns indicate that while the hybrid approach can identify bias presence, fine-grained categorisation remains challenging due to inherent semantic overlap between bias types as explained before in Chapter 3 section 3.2.4.

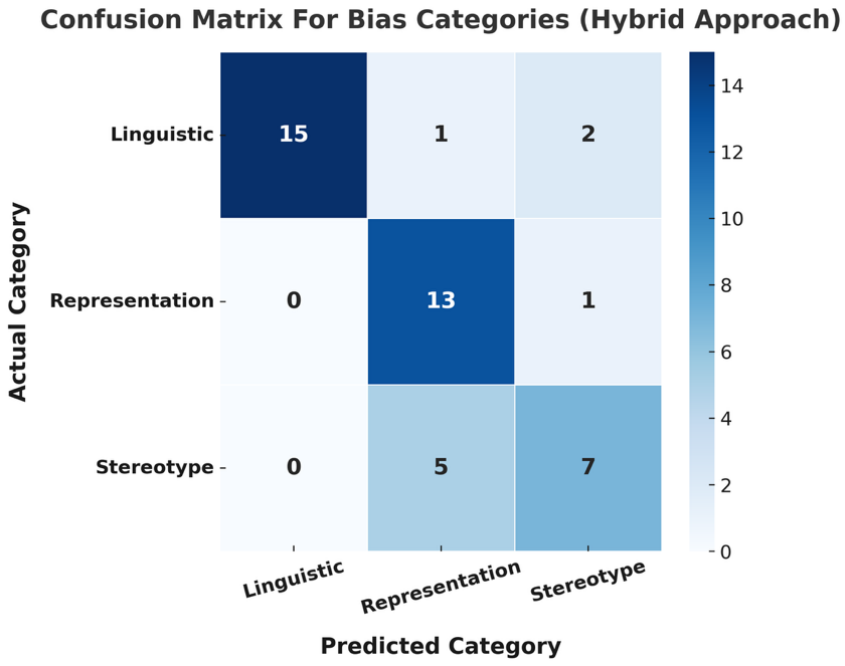

Figure 7.3: Hybrid approach confusion matrix for bias categories

When compared to the fine-tuned LLM-only approach (fine-tuned GPT-4o mini), the hybrid system demonstrates a different error distribution pattern. Whilst the fine-tuned LLM-only model exhibited a tendency to overpredict 'Linguistic' bias, frequently misclassifying 'Stereotype' and 'Representation' instances as 'Linguistic', the hybrid approach shows a shift towards overpredicting 'Representation' bias. This suggests that the hybrid methodology may place greater emphasis on structural and representational elements rather than purely linguistic markers. The hybrid approach achieves relatively consistent performance across bias categories, with correct identifications distributed as 8, 12, and 7 for 'Linguistic', 'Representation' and 'Stereotype' respectively, avoiding the pronounced bias towards a single category that characterised the fine-tuned LLM-only predictions.

The hybrid approach partially addresses the fine-tuned LLM-only system's overprediction of 'Linguistic' bias through integration of multiple analytical components, resulting in more distributed predictions across categories. However, the shift from overpredicting 'Linguistic' to 'Representation' bias indicates that whilst the hybrid system redistributes classification errors, categorical preference persists rather than being eliminated entirely.

While the fine-tuned LLM-only approach (fine-tuned GPT-4o mini) achieved a slightly higher overall accuracy of 81% compared to the hybrid approach's 78%, this difference should be considered alongside other performance factors. The hybrid approach demonstrates more balanced performance across bias classes and categories, which may be beneficial for applications requiring consistent classification across different bias types. However, the choice between approaches depends on whether overall accuracy or distributional balance is prioritised, as each system presents distinct trade-offs in bias detection within academic texts.

7.6 Hybrid Approach Variations

To comprehensively evaluate the contribution of RAG to bias detection, several hybrid system variants were implemented, each representing a distinct approach to combining fine-tuned GPT-4o-mini with RAG-derived precedents. Below is a summary of the main experimental pipelines, their underlying decision logic, and the resulting classification performance.

Each variant was developed to probe specific aspects of hybrid decision-making, from confidence-based overrides to in-context reasoning, with the goal of determining the optimal balance between leveraging annotated precedents and retaining the generative flexibility of the language model. The following section describes each hybrid configuration in detail, elucidating both their operational mechanisms and the reasoning underpinning their design. Table 7.2 below shows a summary hybrid approach variations.

Hybrid V2: Fine-Tuned GPT-4o-mini + RAG (Confidence-Based)
• **Retrieval Method:** Dense vector embeddings (all-mpnet-base-v2) with cosine similarity. • **Thresholds & Logic:** - Fine-tuned GPT-4o-mini confidence > 80%, use GPT-4o-mini prediction - Fine-tuned GPT-4o-mini confidence 60%–80%, use GPT-4o-mini, cross-check with RAG - Fine-tuned GPT-4o-mini confidence < 60%: override with RAG • **Fallback/Override Logic:** If fine-tuned GPT-4o-mini confidence < 60%, prediction is refined/overridden using retrieved example from RAG Corpus • **Accuracy:** 62% • **Macro-F1:** 60%
Hybrid V3: Fine-Tuned GPT-4-mini + RAG (FAISS, Similarity-Based)
• **Retrieval Method:** FAISS with dense vector embeddings (all-mpnet-base-v2) • **Thresholds & Logic:** - Fine-tuned GPT-4o-mini confidence > 0.50: use GPT-4o-mini predictions - Fine-tuned GPT-4o-mini confidence ≤ 0.50 and similarity ≥ 0.25, use RAG annotation • **Fallback/Override Logic:** If confidence ≤ 0.50 and similarity ≥ 0.25, use RAG annotation; else, use fine-tuned GPT-4o-mini prediction • **Accuracy:** 65% • **Macro-F1:** 48%
Hybrid V4: Fine-Tuned GPT-4o-mini + RAG (Prompt-Informed)
• **Retrieval Method:** Dense vector embeddings (all-mpnet-base-v2) with cosine similarity • **Thresholds & Logic:** Always append top-k (i.e., K= 3) retrieved annotated examples from RAG corpus as in-context for fine-tuned GPT-4o-mini prediction • **Fallback/Override Logic:** No fallback/override; output is always from fine-tuned GPT-4o-mini, informed by RAG context • **Accuracy:** 59% • **Macro-F1:** 44%

Table 7.2: Hybrid bias detection system variants and performance

7.6.1 Hybrid V2: Fine-Tuned GPT-4o-mini + RAG (Confidence-Based)

This variant employs a confidence-weighted strategy in which each prediction from the fine-tuned GPT-4o mini model is accompanied by a confidence score. As detailed in the previous chapter, this confidence score is calibrated based on the degree of agreement among human annotators during model training, making it a more reliable indicator of predictive certainty than naïve probability estimates though it is not entirely immune to error.

For each input, if the model's confidence exceeds 80%, the system directly accepts the fine-tuned GPT-4o mini's structured prediction. When the confidence falls within the intermediate range (60–80%), the system still presents the model's output but also retrieves the most relevant annotated example using dense vector embeddings generated by the all-mpnet-base-v2 sentence transformer model, with cosine similarity used to measure semantic closeness between the input text and RAG corpus entries, and substitutes the retrieved annotation for the final output. If the confidence is below 60%, the system relies exclusively on the RAG annotation, fully overriding the model output with the label, category, keywords, and justification from the closest precedent. Importantly, when the system presents a classification derived from a retrieved example whether in the intermediate or low-confidence scenario this is clearly indicated to the user. In such cases, the interface not only shows that the output is informed by precedent but also displays the full text of the RAG example from which the annotation is derived. This approach ensures that users can trace the source of the prediction and understand its context, thereby enhancing transparency and interpretability.

The motivation for this design is to harness the strengths of the model when it is highly confident, while increasing reliability in ambiguous or low-confidence situations by deferring to previously validated human annotations. However, the rigid adoption of RAG annotations even in the intermediate confidence range can introduce abrupt transitions and potential inconsistencies in the user experience. As reflected in the results, this approach achieved an accuracy of 62% and a macro-F1 of 60% but ultimately did not provide the most balanced or interpretable solution for the hybrid bias detection task.

7.6.2 Hybrid V3: Fine-Tuned GPT-4o-mini + RAG (Similarity-Based, FAISS)

In this variant, a similarity-based decision framework is implemented using dense vector retrieval via FAISS (i.e., Facebook AI Similarity Search). For each input, both the user-provided text and the entries in the RAG corpus are encoded as high-dimensional embeddings using the all-mpnet-base-v2 sentence transformer model. FAISS enables efficient nearest-neighbour search across the corpus, allowing the system to quickly identify the most similar annotated examples. Cosine similarity is then computed between the input embedding and the top retrieved examples to rank them by semantic closeness.

The system then applies the following decision rule: if the model's confidence is less than or equal to 0.50 and the cosine similarity with the closest retrieved example is greater than or equal to 0.25, the annotation comprising class, category, keywords, and justification from this precedent is adopted as the final output. Otherwise, the structured prediction generated by the fine-tuned GPT-4o mini model is used.

The reasoning behind this approach is that, even in situations where the model is uncertain, a sufficiently high similarity to a well-annotated precedent provides a robust basis for transferring the RAG annotation. The use of FAISS is beneficial for ensuring scalability and rapid retrieval, particularly as the size of the annotated corpus grows. Several

variations of this decision rule were explored, adjusting both the confidence and similarity thresholds to optimise performance. However, only the best-performing configuration is reported in the summary above.

To maintain transparency, whenever a RAG-based classification is used, the system clearly indicates this to the user, presenting the full text of the precedent from the RAG corpus alongside its annotation. This allows users to see exactly how and why a particular decision was reached, enhancing interpretability and user trust.

Despite these strengths, Hybrid V3 is not without limitations. The relatively low similarity threshold can, at times, result in overrides by precedents that are not truly analogous to the input, increasing the risk of misclassification. Additionally, as with the previous variant, model confidence is not always a perfect proxy for true uncertainty. This approach achieved an accuracy of 65% and a macro-F1 of 48%. While the results demonstrate the feasibility of similarity-based overrides, the limitations noted above led to the selection of an alternative configuration as the final hybrid solution.

7.6.3 Hybrid V4: Fine-Tuned GPT-4o-mini + RAG (Prompt-Informed)

In Hybrid V4, the system uses a prompt-informed strategy, in which the fine-tuned GPT-4o mini model is provided with explicit contextual support from the RAG corpus for every prediction. For each input, the top k (i.e., k= 3) most similar annotated examples are retrieved using dense vector embeddings (all-mpnet-base-v2 sentence transformer) with cosine similarity for ranking. These examples, including their full annotations, are added directly to the prompt given to fine-tuned GPT-4o mini before it generates a prediction.

Crucially, this approach does not employ any fallback or override logic: the output is always produced by the fine-tuned GPT-4o mini model itself. However, both the model and the user are presented with the same three RAG examples. The fine-tuned GPT-4o mini uses these as in-context references to inform its own prediction, and the user sees both the model's structured output (bias class, bias category, bias keywords, and justification) and the full retrieved examples as additional context. In other words, the final output combines the model's generated prediction with the three retrieved RAG examples, so that users can easily see what prior cases may have influenced the model's decision.

This design aims to enhance transparency and support user trust by making the entire reasoning process more explicit. However, this variant achieved an accuracy of 59% and a macro-F1 of 44%, which is lower than the other hybrid approaches. This suggests that while providing explicit context can improve interpretability, it does not necessarily lead to higher predictive performance. The model may not always effectively integrate the retrieved examples into its reasoning, or it may revert to generic responses when the examples are not sufficiently relevant. As a result, this prompt-informed hybrid was not chosen as the final solution.

7.6.4 Limitations and Considerations in Hybrid Variations System Performance

It is important to acknowledge that while the alternative hybrid approaches did not ultimately achieve the highest overall accuracy, their absolute performance should not be viewed in isolation. Several external factors may have influenced the results observed in these experiments. Most notably, the size and diversity of the RAG corpus play a significant role in determining the effectiveness of retrieval-augmented strategies. A relatively limited or unbalanced set of annotated examples may constrain the system's ability to find highly relevant precedents for new inputs, particularly in the case of subtle or less frequently represented bias categories.

Furthermore, the quality and consistency of the annotated corpus itself, the distribution of classes, and the selection of similarity thresholds all impact how well the hybrid decision logic can leverage prior examples. Expanding and continuously updating the RAG corpus with greater coverage across domains, bias types, and linguistic variation could be expected to improve the reliability and utility of these hybrid approaches.

Finally, user-centred considerations such as transparency, interpretability, and the ability to surface relevant precedents remain important strengths of the RAG-based methods, even where raw accuracy may not match that of the most selective or finely tuned variants.

7.7 Limitations and Ethical Considerations

A number of technical, methodological, and data-related limitations should be acknowledged. The RAG corpus, though carefully assembled and partially augmented, is relatively small and heavily influenced by Western educational contexts. As a result, both the retrieval and classification processes may miss or misinterpret bias patterns prevalent in other cultural, temporal, or disciplinary settings. Reliance on LLM-generated annotations, while practical for scaling, raises concerns about annotation consistency and the possible perpetuation of subtle biases.

Methodologically, the use of strict accuracy as a primary metric may underestimate system performance in cases where multiple plausible bias categories exist. The relatively small and imbalanced test set further limits statistical generalisability. On the technical side, the current use of cosine similarity in retrieval, fixed decision thresholds, and static prompt construction may not capture the full complexity of contextual bias detection, particularly for nuanced, intersectional, or indirect forms of bias.

Deploying automated bias detection systems in educational settings has significant ethical implications. There is a risk of over-reliance on automated outputs, potentially leading to inappropriate content flagging or insufficient recognition of subtle, context-dependent bias. Algorithmic biases present in both the training data and the RAG corpus could inadvertently perpetuate or amplify existing inequalities in curriculum curation and content review. The system's design prioritises interpretability by surfacing retrieved precedents

and model justifications, but the complexity of its decision logic may still limit user understanding.

Cultural sensitivity is particularly important, given that the model and corpus are primarily trained on Western sources. Classifications and thresholds deemed appropriate in one context may not translate elsewhere. For this reason, system predictions should always be viewed as decision support for educators, not as substitutes for critical human judgement.

7.8 Conclusion

This chapter has examined the design, implementation, and evaluation of a hybrid bias detection system for educational materials, integrating a fine-tuned GPT-4o mini with RAG. The results demonstrate that the hybrid approach offers clear benefits in interpretability and prediction balance, particularly when compared to standalone language models that may overfit to particular bias categories. By leveraging contextually relevant precedents from a curated corpus, the hybrid system provides more transparent and context-aware justifications for its predictions.

However, the findings also highlight persistent challenges. The most notable difficulties relate to the subjectivity of bias annotation especially in differentiating between explicit and potential bias and the boundaries created by a limited and culturally specific RAG corpus. The analysis further reveals the critical importance of context in educational texts, and the need for systems that can flexibly accommodate ambiguity and multiple perspectives.

Taken together, the work presented in this chapter demonstrates the value and complexity of hybrid approaches for bias detection in academic settings. The hybrid system advances both the technical and ethical dimensions of automated content analysis, emphasising the necessity of transparency, interpretability, and critical human oversight in supporting equitable educational practice. Building on these findings, the following chapter applies the insights and outcomes of this research to the design and test of a web-based application for bias detection in educational content.

Chapter 8

8 Design, Prototyping and Testing of a Web Application for Bias Detection in Online Higher Education Learning Resources

8.1 Introduction

The creation of bias detection applications within higher educational contexts remains a challenging yet necessary extension of theoretical research (Chinta et al., 2024). While extensive studies have demonstrated the prevalence of bias within educational learning materials (Lange and Kelley, 1971; Sadker, D. and Zittleman, 2007; Blumberg, 2008; Bachore and Semela, 2022; Hu, X. and Hancock, 2024), practical tools that allow educators and researchers to identify and mitigate such biases are scarce (Holmes et al., 2019). As part of this research project, a web-based application was designed, prototyped and tested to operationalise the hybrid experiment introduced in the previous chapter. This application provides an accessible solution for detecting and annotating bias, powered by a hybrid system trained on the OER-H dataset.

Building upon the OER-H data collection, annotation methodology (see Chapter 3 section 3.6), LLM model fine-tuning (see Chapter 6 section 6.4) and the hybrid approach (see Chapter 7 section 7.4), this chapter details the practical design of those methods through the prototype of a web application for real-world bias detection in higher education learning materials. This chapter outlines the design requirements, key architectural decisions, system components, model integration process and testing outcomes involved in creating the bias detection web application.

8.1.1 Objectives of the Chapter

The primary objective of this chapter is to build a web application to bridge the gap between advanced AI and NLP methodologies and practical usability for end-users within academic environments. The need for applied AI systems that can address fairness and bias detection in educational resources has been highlighted by recent studies (Holstein et al., 2019; Mehrabi et al., 2021), demonstrating the growing demand for interpretable, actionable AI tools.

8.1.2 Contributions of The Chapter

This chapter aims to fulfil the fourth research contribution outlined in Chapter 1, Section 1.3.4.4, developing a web application prototype for bias detection within higher education content.

8.2 System Requirements

System creation was guided by the specific requirements identified in earlier stages of this research, particularly the classification of bias into clearly defined classes and the categorisation of bias types.

8.2.1 Functional Requirements

The primary functional requirements were defined as follows:

- **Bias class detection**: The system must classify input excerpts accurately into one of three classes: 'Bias', 'Not Bias' or 'Potential Bias'. This structure mirrors the classification schema developed during the earlier data annotation outlined in Chapter 3 section 3.6.
- **Bias category identification**: For text classified as Bias or Potential Bias, the system must further categorise the detected bias into one of three specific categories: 'Stereotype', 'Representation' or 'Linguistic'. This categorisation provides users with detailed insights into the type of bias present, thereby enhancing the interpretability of the system outputs.
- **Bias keywords/phrase extraction**: Where bias is detected, the system must also printout the specific words or phrases within the input text that contributed to the classification, which may not necessarily be explicitly biased words but could still play a role in shaping the biased interpretation.
- **Bias decision justification**: In order to support explainability and user trust, the system must generate an explanation for each classification decision.
- **User feedback collection**: The system must enable users to respond to each classification result by selecting either agree 'Yes' or disagree 'No'. This feedback mechanism supports human-in-the-loop learning and is designed to inform future iterations of model refinement. Collecting the binary agreement data aligns with user-centred AI development principles and has been shown to improve robustness and accountability in deployed systems (Amershi et al., 2019).

8.2.2 Non-Functional Requirements

The non-functional requirements defined for the system included:

- Privacy and data security: User-submitted texts and feedback are stored to support system evaluation and research. Users are explicitly informed, prior to using the tool, that their input will be retained. All data is transmitted over encrypted channels using Transport Layer Security (TLS), to ensure secure communication between the Chrome extension app and the backend service. No personally identifiable information is collected, and all stored content is used solely for research purposes, in accordance with ethical guidelines for AI in education (Williamson and Eynon, 2020).

- Responsiveness: To ensure practical usability, the system must deliver classification results within a maximum of five seconds per excerpt.
- User guidance: The app must include a clear, accessible interface with basic usage instructions and an example input of different bias scenarios. This improves usability by supporting first-time users and reducing barriers to adoption in non-technical academic settings.
- Browser-based accessibility: Given the widespread usage of Chrome among academic institutions (StatCounter, 2023), the system was created as a Chrome extension to facilitate immediate and easy user adoption without additional software installations.

8.3 System Architecture

The bias detection system is structured as a client–server architecture, in which the Chrome extension operates as the client and communicates with a backend service implemented using FastAPI (Ramírez, 2018). This architectural separation supports independent development and maintenance of the user interface and server-side logic. All communication is conducted over encrypted channels using TLS. The design facilitates scalability, integration with external services, and future system extension. The main components include the Chrome extension (frontend), the FastAPI backend, the fine-tuned GPT-4o mini accessed via the OpenAI API, RAG, and a storage layer for storing user inputs and feedback (see Figure 8.1).

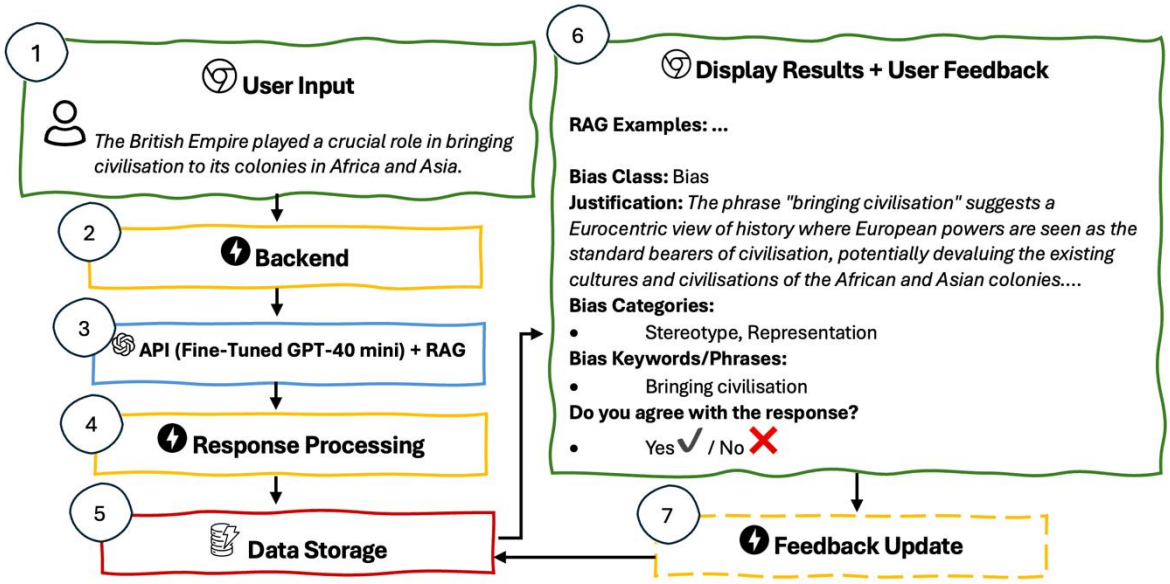

Figure 8.1: System architecture for the web application for bias detection in online higher educational learning materials

8.3.1 Frontend: Chrome Extension

The Chrome extension provides a browser-based interface through which users can submit text for analysis. It allows users to:

- View an introductory screen with usage instructions and example inputs.
- Enter input text for analysis.
- Submit the text to the backend for processing.
- View the classification results, including the bias class (Bias, Not Bias, or Potential Bias), bias category (if applicable), justification and extracted biased keywords or phrases.
- Provide feedback on the classification output by selecting "Yes" (agree) or "No" (disagree).

8.3.2 Backend API: FastAPI

The backend service is built using FastAPI (Ramírez, 2018), which handles the client requests. It performs the following tasks:

- Receives and validates user-submitted text.
- Forwards the input for classification.
- Invokes the RAG, if required.
- Returns the structured output to the frontend.
- Logs the input, prediction results, and any user feedback.

The API is designed to support low-latency responses through asynchronous request handling and can be extended or modified independently of the frontend component, ensuring modularity and scalability.

8.3.3 Inference Pipeline with OpenAI API and RAG

8.3.3.1 Interaction with the Fine-Tuned Model

The central component of the pipeline is a fine-tuned GPT-4o mini model, accessed through the OpenAI API. For each input sentence requiring classification, the system performs the following:

- Prompt Construction:
  A structured prompt is generated, containing:
  - The input sentence to be analysed for bias.
  - The k most similar annotated examples (typically, k=3), drawn from retrieval corpus (i.e., RAG corpus), each with its bias class, category, keywords and justification.
  - Explicit instructions directing the model to output its answer in a structured JSON format containing:

- A bias class (Bias, Potential Bias, or Not Bias).
- One or more bias categories (where applicable).
- Associated bias keywords (where applicable).
- A concise justification.
- A confidence score representing the model's self-assessed certainty.

- Model Response and Output Handling

  The model's output is expected as a JSON string. To ensure robustness, all outputs are passed through a dedicated repair-and-parse function, which corrects common formatting errors (e.g., mismatched quotes, trailing commas, or partial outputs), guaranteeing reliable downstream parsing and evaluation.

8.3.3.2 The Role of RAG: How the Hybrid System Works

The final bias prediction for each sentence is always produced by the fine-tuned model. However, this prediction is explicitly informed by the context of retrieved annotated examples provided as part of the prompt. The model thus synthesises information from both the new input and relevant precedents, using the latter as a form of concrete evidence to guide its decision.

To illustrate, suppose the input sentence is:

> "This text explores whether certain groups possess inherent qualities that justify their societal roles."

- The system embeds this input and retrieves the three most similar examples from RAG corpus; let's assume that all three were previously labelled as Bias due to stereotyping language. The resulting prompt shown to the model appears as follows:

> **System instruction:** "You are an expert bias detector..." (Full prompt can be found in Chapter 6 section 6.4)
>
> **Input educational text:** This text explores whether certain groups possess inherent qualities that justify their societal roles.
>
> **Similar annotated examples retrieved:**
>
> 1. [Example text 1] 70% similarity
>   - Label: Bias
>   - Category: Stereotyping and Representation
>   - Justification: [text]
> 2. [Example text 2] 60% similarity
>   - Label: Bias

> - Category: Language and Tone
> - Justification: [text]
>
> 3. [Example text 3] 45% similarity
>   - Label: Potential Bias
>   - Category: Stereotyping and Representation
>   - Justification: [text]
>
> **Model task instruction:** Using the above, classify the input as Bias, Not Bias, or Potential Bias.
> Provide category, keywords, and a brief justification.
> Respond in this JSON format: ["result":..., "category":..., "keywords":..., "justification":...]

The model then generates its prediction, e.g., 'Bias' with 'Stereotype' and 'Representation' as category and an explanatory justification, having considered both the input and the retrieved examples.

Thus, the RAG retrieval process does not itself make the classification but provides context and evidence that inform the fine-tuned model's final decision. This approach allows the system to combine the strengths of precedent-based reasoning (as in case-based legal or educational decisions) with the adaptive generalisation of an LLM.

The final classifications are presented alongside relevant examples retrieved from the RAG corpus, each showing a similarity percentage to provide users with greater transparency into the decision-making process. These RAG-derived examples include only the essential classification elements: a brief text excerpt to illustrate the type of content retrieved, the bias class, bias category, and justification. Bias keywords are intentionally excluded to optimise space usage within the Chrome extension interface whilst maintaining clarity and usability.

8.3.4 Response Processing

Once the final classification is obtained the backend prepares a structured response. This response includes:

- The bias class (Bias, Not Bias or Potential Bias).
- The bias category (if applicable).
- A justification explaining the classification.
- Extracted biased phrases within the submitted text.
- Model confidence score.
- Any retrieved example from RAG corpus used in the decision.

This output is formatted as a structured JSON object to ensure consistency and compatibility with the frontend. It serves as the basis for both display and storage operations.

Processing also includes any necessary sanitisation or formatting (e.g., sentence truncation, token de-highlighting) to prepare the output for user-facing display.

8.3.5 Data Storage

Following response processing, the structured classification data along with the original input and any retrieved content is stored in Amazon Web Services (AWS) DynamoDB[24], a fully managed NoSQL database service. The storage layer records:

- Submitted user text.
- Classification decision and explanation.
- Retrieved content (if applicable).
- Model confidence score.
- Timestamp and internal versioning metadata.

This stored information supports future system evaluation, bias case analysis, and potential model retraining.

8.3.6 Content Display (Frontend Rendering)

Once the response is received by the Chrome extension, it is parsed and rendered for user interaction. The frontend displays:

- The bias classification prediction.
- Retrieved examples (if included).
- The feedback buttons 'Yes' or 'No'.

The display layout is designed to be visually clear and non-intrusive, using accessible colour coding and icons where necessary. The results are presented in a collapsible interface that integrates seamlessly with the user's browsing environment.

8.3.7 Feedback Update

Finally, the system enables users to provide feedback on the classification. This occurs via an embedded response form that allows users to select if they agree or disagree with the system's prediction.

This feedback is sent to the backend and linked with the original classification and timestamp. It is stored alongside the classification data and is intended to support ongoing system refinement and evaluation. The feedback mechanism is optional.

[24] https://aws.amazon.com/dynamodb/

8.4 Frontend Implementation

The frontend component of the system was implemented as a Chrome browser extension to ensure accessibility and ease of use within users' existing web-based reading environments. This section describes the user interface design decisions and the visual presentation of the classification outputs.

8.4.1 User Interface (UI) Design Decisions

The interface was designed to be accessible, and minimally disruptive to the user's browsing experience. As shown in Figure 8.2, 8.3, 8.4, 8.5 and 8.6 the extension provides a clear and focused input area where users can manually enter a text for analysis. Above the input field, the interface presents a disclaimer and short description of the system's purpose, scope, and domain coverage. This includes a note that the system was trained on materials from the humanities subjects, with a particular emphasis on detecting gender and ethnic bias in these materials.

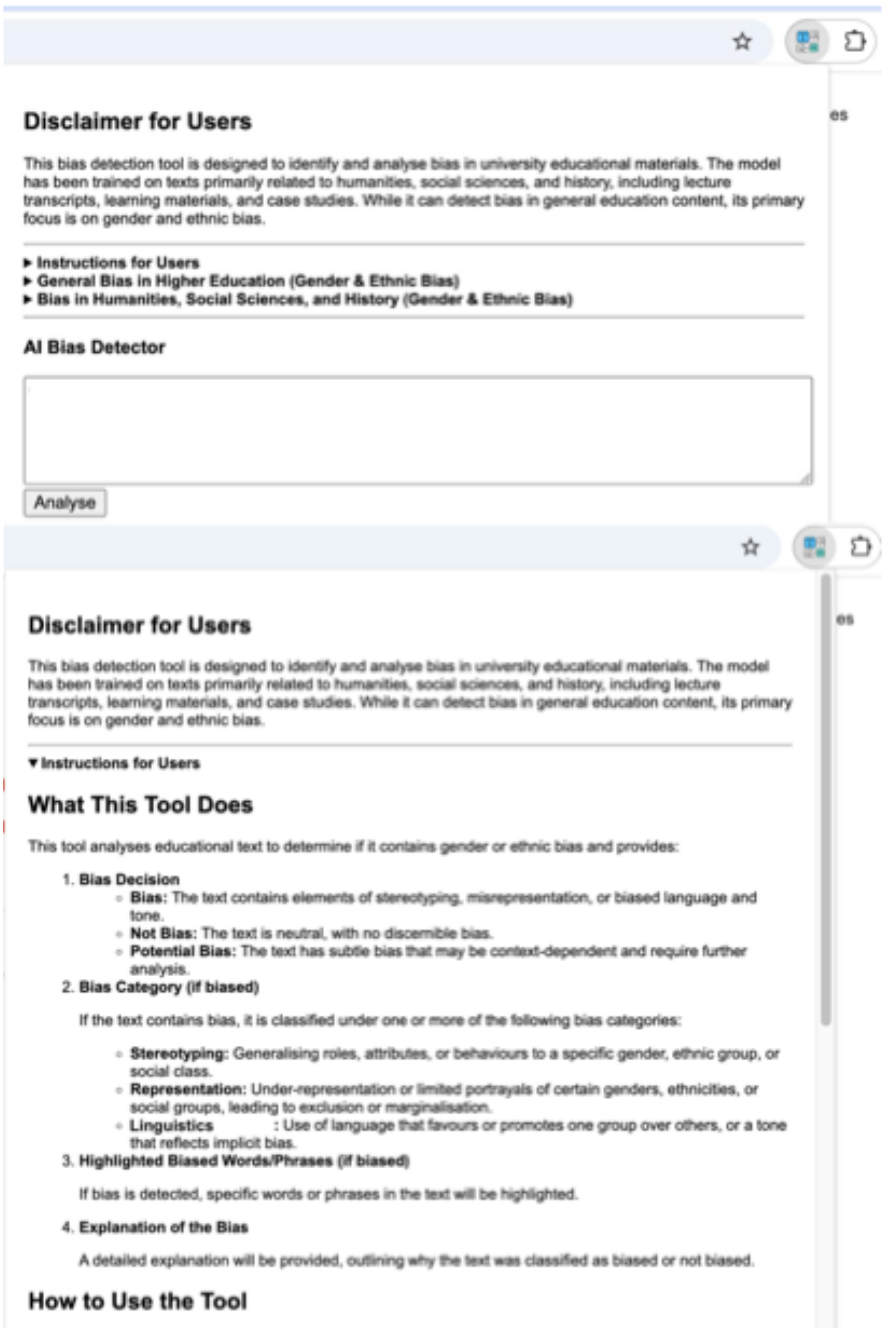

Figure 8. 2: Bias detection application interface with user instructions and general and domain-specific examples of gender and ethnic bias in higher education – Snippet 1

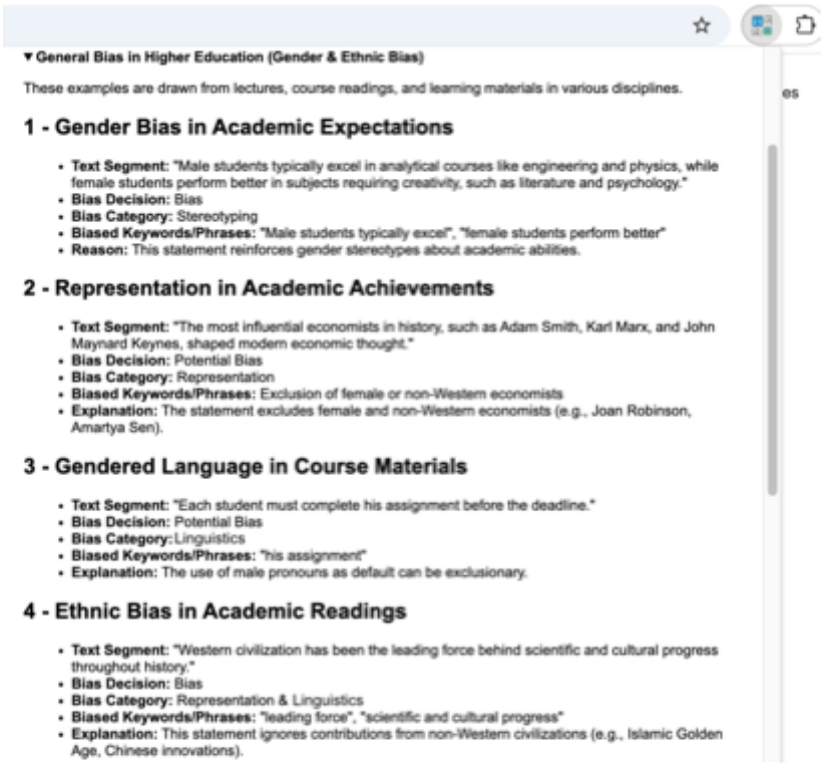

Figure 8. 3: Bias detection application interface with user instructions and general and domain-specific examples of gender and ethnic bias in higher education – Snippet 2

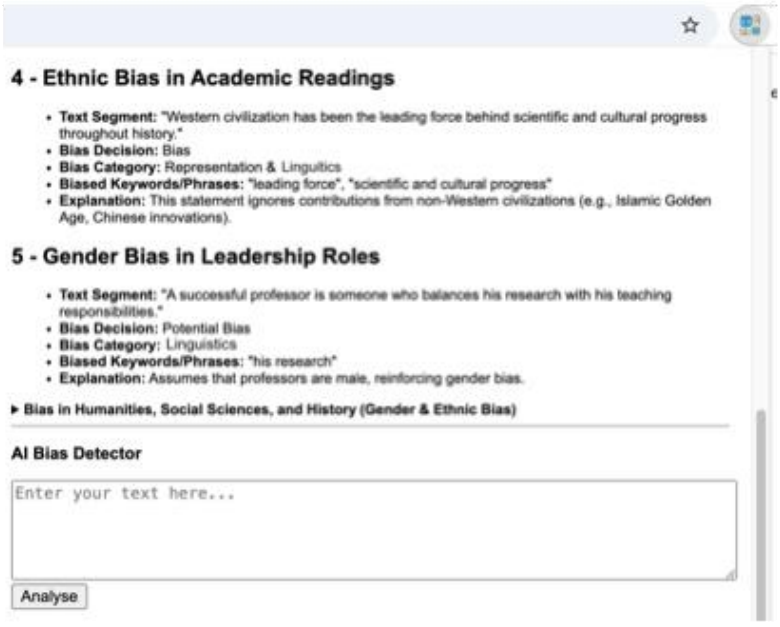

Figure 8. 4: Bias detection application interface with user instructions and general and domain-specific examples of gender and ethnic bias in higher education – Snippet 3

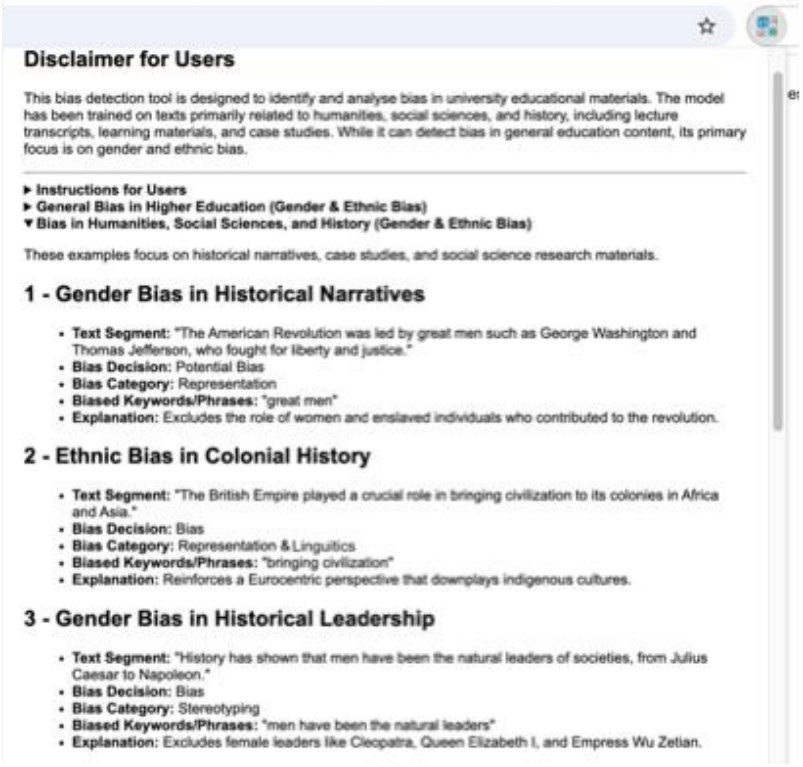

Figure 8. 5:Bias detection application interface with user instructions and general and domain-specific examples of gender and ethnic bias in higher education – Snippet 4

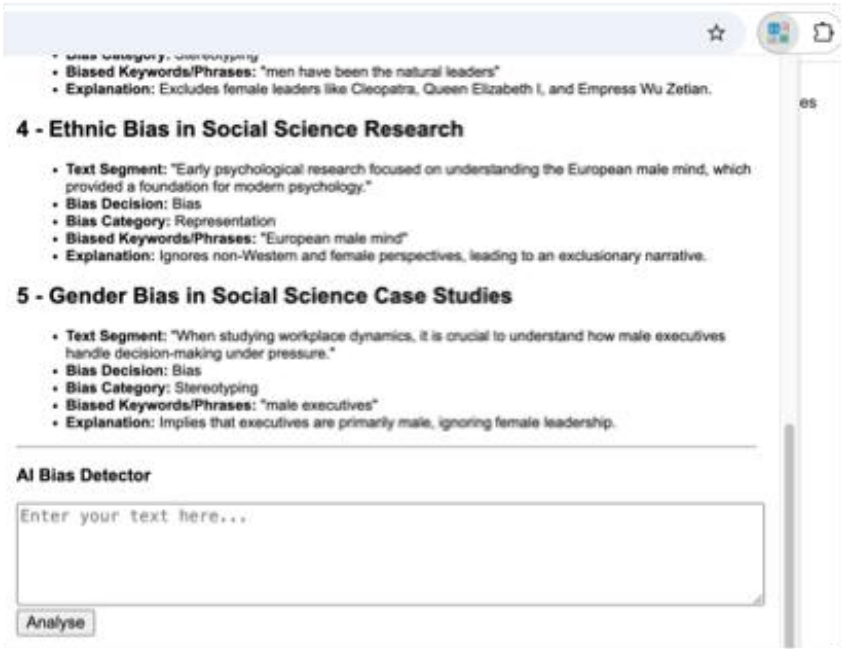

Figure 8. 6: Bias detection application interface with user instructions and general and domain-specific examples of gender and ethnic bias in higher education – Snippet 5

Expandable sections titled “Instructions for Users”, “General Bias in Higher Education”, and “Bias in Humanities” are included to provide optional guidance and context of gender and ethnic bias in both general and domine specific subjects. These elements were designed to support first-time users and ensure that the tool’s function and limitations are clearly communicated.

The UI is intentionally simple, with a single button ‘Analyse’ to submit the input for classification. Upon processing, the results are displayed directly below the input area in the same panel.

After submitting a text for analysis, the user is presented with a structured classification result within the extension interface (see Figure 8.7). The output includes the bias classification prediction, the retrieved examples (if included) from the RAG corpus along with their similarity percentages for added transparency. At the bottom of the result panel, the system presents a binary feedback prompt: “Do you agree with the response?”, with ‘Yes’ and ‘No’ options.

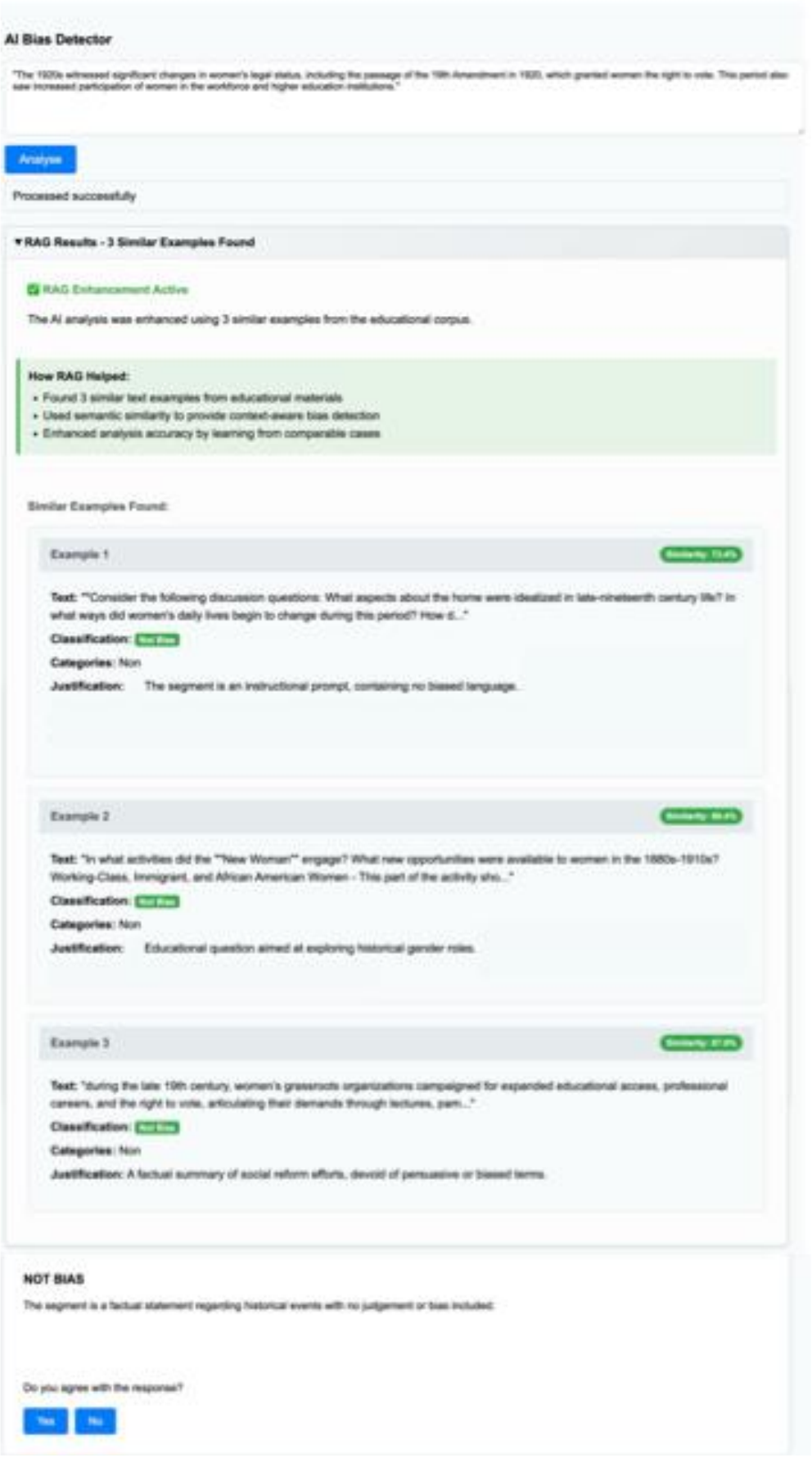

Figure 8. 7: Classification results screen showing bias prediction, retrieved RAG examples and user feedback options

8.4.2 Interface Display of System Classifications: Examples Across Different Bias

To illustrate the practical functionality of the deployed bias detection system, this section presents examples that been tested to show how classification results are displayed within the Chrome extension interface. The following examples demonstrate the system's ability to detect different types of bias across various academic text, showcasing both the technical capabilities and user-facing interpretability features of the application.

8.4.2.1 Not Bias Classification Example

The system demonstrated effective classification capabilities when processing a critical statement about American federal government actions during the Reconstruction era (1865-1877): "The federal government only acted on civil rights when it was politically convenient, but out of any genuine concern for justice." This text evaluates federal policy during the post-Civil War period when the government was tasked with rebuilding the South and establishing civil rights for formerly enslaved people. The fine-tuned model classified this as 'Not Bias' with a confidence score of 57.6%, generating a justification that the text presents "a critical view of the federal government's actions during the period" but constitutes "valid political consideration rather than using language that is inappropriate or discriminatory."

The RAG component retrieved three supporting examples with similarity scores ranging from 51% to 57%, all classified as 'Not Bias'. These examples included historical analyses such as: "But white conservatives chafed at the influx of black residents and the establishment of biracial politics" and "The federal government responded to southern paramilitary tactics by passing the Enforcement Acts between 1870 and 1871." The consistency across all three RAG examples reinforced the model's classification decision.

The system's performance illustrates how the RAG enhancement functions as a validation mechanism, where retrieved examples with moderate similarity scores collectively influenced the final output. Despite the relatively modest similarity percentages, the unanimous 'Not Bias' classification across all retrieved examples provided sufficient precedent for the system to maintain its initial assessment. This demonstrates the model's ability to distinguish between critical commentary and biased language through pattern matching with similar textual contexts, showcasing the effectiveness of combining fine-tuned classification with retrieval-based validation.

8.4.2.2 Potential Bias Classification Example

The application shows capability in identifying subtle bias through its 'Potential Bias' classification. When analysing the text "Lincoln may have appointed Davis a general, but let's not pretend his promotion wasn't influenced by politics more than merit," concerning Civil War-era military appointments, the system identified potential bias with a confidence

score of 56.3%. The classification included both 'Linguistic' and 'Representation' bias categories, with highlighted bias keywords including "let's not pretend" and "merit".

The justification explained that the text "reveals a tone of scepticism regarding Davis's promotion, suggesting that it may have been influenced more by political factors than merit. This could be interpreted as implying that Davis did not truly earn his position, casting doubt on his abilities or military achievements". The RAG enhancement retrieved three similar examples, all dealing with historical military appointments and political influence. These retrieved examples demonstrated patterns of language that question individual merit based on political considerations, providing contextual support for the potential bias classification.

The first retrieved example (similarity: 66.3%) stated: "It is important to understand, however, that the same historically creative tensions that marked events such as the Freedom Rides played an important role in the March's organisation, program, and imp...". This example was originally classified as 'Not Bias' but demonstrated how historical analysis can navigate complex political motivations without casting aspersions on individual merit. The second retrieved example (similarity: 54.8%) discussed "The first part of this lesson focuses on the Freedom Riders. It demonstrates the critical role of activists in pushing the Kennedy Administration to face the contradiction between its ideals and the..." which was also classified as 'Not Bias' and showed balanced analysis of political pressure and decision-making. The third example (similarity: 51.0%) analysed "JFK, Freedom Riders, and the Civil Rights Movement…" and was classified as 'Not Bias', providing historical context about political appointments during periods of social change.

Notably, all three retrieved examples were classified as 'Not Bias', yet the model successfully identified the input text as 'Potential Bias'. This demonstrates the system's sophisticated understanding of context whilst the retrieved examples use neutral language to discuss political factors in decision making the text employs more sceptical and dismissive language ('let's not pretend') that crosses into potential bias territory by directly questioning merit rather than simply analysing political factors. The RAG enhancement thus enabled the system to make nuanced distinctions between legitimate historical analysis and language that potentially undermines individual achievements through political considerations.

8.4.2.3 Clear Bias Classification Example

When processing text containing explicit bias, the system provides comprehensive analysis with multiple supporting elements. For the input "The so-called 'moral leadership' of women was just a clever excuse to keep them locked in the home and out of real power", the system identified 'Bias' with a confidence score of 61.2%. The bias categorises included both 'Stereotype' and 'Linguistics', highlighting key bias phrases such as "moral leadership", "clever excuse", "real power", "keep them locked in the home" and "out of real power".

The justification explained that the text "contains a biased statement that portrays women's aspirations for leadership and equality as mere manipulation or an excuse rather than a valid desire for societal change. The term 'clever excuse' dismisses the genuine goals of these women, introducing a negative tone that undermines their efforts. This framing reflects bias against acknowledging or valuing women's leadership and introduces traditional gender roles".

The retrieved RAG examples were relevant historical texts dealing with women's suffrage and gender roles. Each retrieved example was classified as 'Bias' under similar bias categories, with justifications explaining how dismissive language towards women's political participation reinforces harmful stereotypes. One retrieved example discussed women's grassroots organisations and was classified as bias due to "persuasive or biased terms", whilst another analysed nineteenth-century attitudes towards women's education, highlighting how historical accounts can perpetuate gender stereotypes.

8.5 Backend Implementation

The backend of the bias detection application is developed using FastAPI, a high-performance, asynchronous web framework in Python well-suited for building robust APIs. It serves as the central processing component of the system, responsible for coordinating communication between the frontend and various backend services. These include interfacing with the fine-tuned GPT-4o mini model via the OpenAI API, executing retrieval operations from RAG corpus, and managing the secure storage of classification results and user feedback in AWS DynamoDB the cloud-hosted database. This section details the design of the backend architecture, with particular focus on its API endpoints, request handling and security mechanisms, and the integration of the RAG corpus.

8.5.1 API Endpoints

The backend exposes a set of Hypertext Transfer Protocol Secure (HTTPS) endpoints that facilitate communication between the Chrome extension frontend and the model-serving logic. The primary endpoints include:

- **/classify**: Accepts a POST request containing the input text. It returns a structured JSON response including the predicted bias class, justification, confidence score, bias category and extracted bias phrases (if applicable). Any retrieved content (i.e., RAG corpus examples) used during inference.
- **/feedback:** Accepts user feedback submitted from the frontend, including their feedback of agreement or disagreement. This endpoint also receives the associated input text and classification output at the time the feedback was given. All of this information is stored together in the database, enabling alignment between user responses and specific system decisions for future evaluation and model improvement.

These endpoints are documented internally using FastAPI's automatic schema generation, allowing for easier testing and potential extension in future iterations.

8.5.2 Request Handling and Security

All client-server interactions are secured using TLS-encrypted HTTPS communication to ensure data confidentiality in transit. Each incoming request is validated to ensure proper format and completeness before further processing. Sensitive actions such as database writes are protected with API key-based authentication to prevent unauthorised access, particularly in hosted or production environments.

Input text is pre-processed to remove empty or malformed inputs. Exception handling is implemented to gracefully manage timeouts, API failures, or retrieval errors, ensuring a robust user experience and reducing the likelihood of interface crashes.

8.5.3 Retrieval from RAG Corpus

For inputs where the model's confidence is below or equal to a predefined threshold (70%), the backend invokes a retrieval module that searches a static corpus of annotated bias examples using dense vector embeddings generated by the all-mpnet-base-v2 sentence transformer model. The corpus is stored in a lightweight local index, and retrieval is based on cosine similarity between the vectorised input and stored excerpts. The top K (i.e., K=3) most similar annotated examples are retrieved and included directly in the prompt provided to the fine-tuned GPT-4o mini model. The model uses these retrieved examples as in-context references to inform its classification decision. This retrieval process is integrated into the main classification logic and is automatically triggered based on the confidence score returned by the fine-tuned GPT-4o mini model.

8.5.4 Feedback Mechanism

The system includes a feedback mechanism that is designed to serves a dual purpose: it encourages user engagement with the bias detection process, and it supports the long-term improvement of the system through user-informed evaluation and potential retraining.

After submitting a text excerpt for analysis and receiving the classification output, users are prompted with a simple binary question: "Do you agree with the response?". They may respond by selecting either Yes or No. The feedback is transmitted to the backend along with the original input text and the corresponding classification output. This information is stored in AWS DynamoDB. Each feedback entry is timestamped and stored as a unique record, enabling longitudinal tracking of user agreement trends and identification of recurring disagreement patterns. The database structure is designed to associate feedback not just with the user interaction, but with the exact decision made by the system at that point in time.

8.6 Application Evaluation – Pilot Study and Testing

This section presents the evaluation of the web-based bias detection application practicality and functionality. While previous chapters have addressed the underlying model's

performance using annotated datasets and comparative baselines, the present evaluation seeks to move beyond algorithmic accuracy to examine the operational effectiveness of the system as an accessible educational tool. This evaluation encompasses both internal testing conducted by our research team to assess system functionality and reliability, as well as a pilot study with users to evaluate the system's usability and practical application in educational contexts. The pilot study is placed before internal testing because it provides initial user feedback that informs the more detailed internal evaluation.

8.6.1 Pilot Study

8.6.1.1 Participants

The usability pilot study involved two participants, both of whom are postgraduate students in computer science and are familiar with higher education academic materials and research workflows.

- **Participant 1:** A graduate student currently employed as a software developer in a medical research context. This participant holds a master's degree in advanced computer science (artificial intelligence) and regularly engages with technical and academic texts as part of their professional responsibilities.

- **Participant 2:** A PhD student specialising in computational linguistics, with experience as a university lecturer. This participant holds a master's degree in advanced computer science (data analytics) and possesses practical teaching experience as well as extensive knowledge of ML frameworks, scholarly publication in AI, and NLP techniques.

Both participants were selected on the basis of their background in academic research and their familiarity with evaluating educational content and web-based applications. Their selection was facilitated through existing academic and professional networks.

8.6.1.2 Tasks

A pilot study was conducted to evaluate the usability and effectiveness of the bias detection application. The study utilised a set of carefully curated text examples drawn from higher educational materials in Humanities disciplines, designed to represent varying degrees and types of bias. These examples were human-annotated to establish ground truth classifications. Eight test examples were organised into four categories: Bias Cases (Examples A1-A2), Potential Bias Cases (Examples B1-B2), Not Bias Cases (Examples C1-C2), and Complex/Ambiguous Cases (Examples D1-D2). The final two examples (D1 and D2) were deliberately included as they had demonstrated lower inter-annotator agreement during the human annotation process, reflecting genuine ambiguity in bias detection tasks. Participant 1 evaluating examples A1, B1, C1, and D1, whilst Participant 2 assessed examples A2, B2, C2, and D2.

8.6.1.3 Data Collection

During the pilot sessions the data was collected through a set of structured questions including:

- What aspects of the interface or workflow were clear or confusing?
- Did you find the bias classification and justifications understandable and useful?
- Would you consider using this tool in your academic or professional work?
- How easy was it to use the application as an extension?
- Did you encounter any technical issues or unexpected behaviour?

The evaluation relies on qualitative user feedback, a well-established approach in applied NLP research for assessing automated text analysis tools (Crowston et al., 2012; Leeson et al., 2019; Abram et al., 2020). These insights help refine the tool and support its use in academic workflows.

8.6.1.4 Pilot Study Outcomes and User Insights

1. User Experience

Both participants found the interface and instructions to be clear, indicating successful initial user engagement with the application. The process of inputting text for analysis was deemed intuitive by both participants, suggesting the interface design effectively supports user workflow. Moreover, participants found the feedback mechanism very easy to use, indicating successful implementation of user interaction features.

2. Classification Results

Participants generally found the bias classifications comprehensible and useful. However, an interesting phenomenon emerged regarding the justification outputs that warrants methodological discussion. Occasionally, explanations included prefatory labels such as 'AI' and 'Annotators' before providing the actual reasoning for the bias classification. For instance, outputs would display: "Annotators: no bias language is used and people answering these questions...etc" or "AI: This text is biased as it perpetuates a Eurocentric view of the development of modern democracy...etc". Whilst participants noted this occurred infrequently, it provides valuable insight into how models internalise training data formatting.

This phenomenon can be traced directly to the data annotation methodology detailed in Chapter 3. All data used for fine-tuning were annotated by human annotators, who built upon initial AI-generated annotations as a starting point. In cases where human annotators agreed with the AI's initial classification but provided different explanatory reasoning, both the original AI justification and the human annotator's explanation were retained in the training data. This approach was intentionally designed to preserve diverse reasoning

styles and expose the fine-tuned model to multiple explanation frameworks. Additionally, retaining both sets of explanations allowed us to examine how different human annotators were perceiving bias compared to the AI system, revealing fascinating differences in reasoning patterns that provided valuable insights into the human annotation process, and the varied cognitive approaches humans employ when identifying and justifying bias classifications.

The observation that the model occasionally learned these labels ('AI:' and 'Annotators:') as formulaic prefixes represents a noteworthy finding about how models process mixed training data sources. Rather than understanding these labels as metadata indicators distinguishing the source of different reasoning approaches, the model appears to have internalised them as part of the explanatory structure itself. This suggests an interesting tension between preserving methodological transparency in training data and ensuring clean model outputs.

This finding contributes to our understanding of fine-tuning methodologies in several ways. Firstly, it demonstrates that whilst incorporating both AI and human reasoning patterns successfully enhanced the model's explanatory capabilities, training data formatting can inadvertently influence output structure. Secondly, it highlights the importance of considering how metadata and source indicators in training data may be interpreted by models during the learning process. These insights have implications for future work in human-AI collaborative annotation and suggest that careful consideration of data formatting protocols is essential when developing explanation-generating systems.

The identification of this pattern also enabled the development of targeted post-processing refinements to ensure consistent user experience, demonstrating the value of comprehensive evaluation methodologies in surfacing such phenomena.

3. Comments and Technical Issues

Participants provided positive feedback regarding the overall clarity of the interface, with one noting: "The interface and instructions were clear." However, technical inconsistencies were highlighted, particularly regarding incomplete outputs where "sometimes the output isn't complete (missing bias keyword or bias category)."

A notable pattern emerged where participants encountered incomplete outputs during initial analysis attempts. Both participants discovered that clicking 'analyse again resolved this issue, producing complete outputs whilst maintaining consistency with the previous bias classification decision. This suggests a technical reliability issue that does not affect the core analytical functionality but impacts user experience.

4. Professional Application Consideration

When asked whether they would consider using this tool in their academic or professional work, participants responded positively with 'Yes/Maybe' suggesting the application

demonstrates sufficient utility and reliability for professional contexts, pending resolution of the identified technical issues.

The pilot study results indicate that whilst the application successfully delivers its core functionality with an intuitive interface, minor technical refinements are needed to ensure consistent output completion and clearer justification formatting.

8.6.1.5 Pilot Study Discussion and Limitations

The pilot study results demonstrate that the bias detection application successfully achieves its primary objective of providing accessible and interpretable bias analysis for educational content. Both participants found the interface intuitive and the classification results generally understandable, indicating effective user experience design. The technical issue of incomplete outputs, whilst concerning, appears to be a display formatting problem rather than a fundamental analytical flaw, as re-running analyses consistently produced complete results without altering the bias classification. This suggests the core machine learning model functions reliably whilst the user interface requires minor refinements.

The appearance of 'AI:' and 'Annotators:' prefixes in justifications reveals an interesting artefact of the training methodology, illustrating how the model learned formatting patterns from the dual-annotation training data where both AI-generated and human explanations were preserved. Whilst this created occasional confusion, it also demonstrates the model's ability to incorporate diverse reasoning approaches, potentially enriching the explanatory content when properly formatted.

However, several limitations must be acknowledged. The pilot study involved only two participants, which severely limits the generalisability of findings. This sample size is insufficient to identify patterns across different user demographics, educational backgrounds, or technical proficiencies that might significantly impact user experience and application effectiveness. Participants may not adequately represent the diverse user base expected to engage with the application, including educators from various disciplines, students at different academic levels, and professionals with varying degrees of bias awareness and technical expertise. Cultural and linguistic diversity considerations were not addressed in this limited sample.

Furthermore, no formal accessibility testing was conducted beyond basic observational notes. The application's usability for users with disabilities, varying technical capabilities, or different learning preferences remains unexplored, potentially excluding significant portions of the intended user base. The model's core evaluation continues to rely on the original human-annotated dataset rather than live, user-generated content, meaning the system's performance on real-world educational materials remains largely untested. The pilot examples, whilst carefully curated, represent a narrow range of bias types and academic disciplines, and the incomplete output problem requires more systematic

investigation to understand its frequency, triggers, and potential impact on user trust and adoption.

8.6.2 Internal Testing and System Validation

8.6.2.1 Technical Implementation and Testing

The bias detection application was implemented as a Chrome extension with backend services hosted on a MacBook Pro (Apple M2 Pro, 32GB RAM). The architecture comprised a FastAPI backend, RAG pipeline, and AWS DynamoDB storage, with TLS-encrypted communication and secure OpenAI API integration.

The system achieved median response times of 10 seconds for initial submissions and 2 seconds for subsequent analyses. Initial processing delays (8-12 seconds) occurred due to RAG corpus loading, whilst subsequent responses remained within the five-second target threshold.

No critical failures occurred during testing. However, both participants encountered intermittent issues requiring resubmission in approximately 20-25% of cases, where incomplete results lacked bias categories or keywords. Upon resubmission, complete outputs were generated with consistent core classifications, indicating stable underlying detection logic despite display-related inconsistencies.

Spot-checks confirmed that complete results matched previously validated model performance, with technical issues appearing related to output parsing rather than model inconsistencies.

8.6.2.2 Testing with Standard Examples

The same examples provided to pilot study participants were tested to assess output consistency and classification performance across different bias types. Complete outputs are presented in Tables 8.1-8.8.

The application demonstrated strong alignment with human annotations in most cases. Clear bias examples (A1, A2) were correctly identified as 'Bias', matching human recognition of gender stereotyping and Eurocentric representation. Neutral content (C1, C2) was accurately classified as 'Not Bias', showing the model's ability to distinguish factual accounts from biased material.

For 'Potential Bias' cases (B1, B2), the system correctly identified 'Potential Bias' classifications, matching human annotators in recognising subtle language bias and complex interplays of stereotype and representation biases.

The most revealing results emerged from deliberately ambiguous examples (D1, D2), where human annotators showed lower agreement, resulting in dual ground truth labels ('Potential Bias' and 'Not Bias'). The application classified both as 'Not Bias', aligning

with one valid interpretation in each case. For D1 (Underground Railroad), the system provided balanced justification acknowledging both heroic resistance and historical oppression. For D2 (pre-industrial families), it rationally argued the text presented factual patterns without value judgements.

Retrieved examples consistently supported clear bias classifications with high similarity scores (e.g., 72%). However, some lower scores (e.g., 47.1%) indicated opportunities for corpus refinement.

Aspect	Ground-truth Classification	Bias Detection Application Classification
Text Segment	"Throughout history, women have naturally excelled in nurturing roles such as teaching and nursing, whilst men have demonstrated superior capabilities in leadership positions and scientific endeavours. This biological predisposition explains why most great philosophers and scientists have been male."	-
Bias Class	Bias	Bias
Bias Category	Stereotype	Stereotype
Bias Keywords/Phrases	"naturally excelled," "nurturing roles," "superior capabilities," "biological predisposition," "great philosophers and scientists have been male"	"women have naturally excelled in nurturing roles such as teaching and nursing"
Justification	Contains clear gender stereotypes, biological determinism, and reinforces traditional gender roles.	Stereotyping of inherent gender abilities (women nurturing, men superior in leadership and science) suggests fixed roles and underrepresentation of women in key fields.
Retrieved Examples from RAG Corpus	-	**- Retrieved Example 1 (similarity score: 72%):** **Text Segment:** "Throughout the mid-20th century, biology textbooks frequently depicted women as assistants in the laboratory, relegating them to roles of note-taking and specimen preparation while men conducted the ..." **Bias Class:** Bias **Bias Category:** Stereotype, Representation **Justification:** This segment attributes fixed lab roles to women (stereotyping bias) and underrepresents their scientific contributions (representational bias), reinforcing a narrative that science is primarily a male pursuit. **- Retrieved Example 2 (similarity score: 63.9%):** **Text Segment:** "Modern biomedical research laboratories have seen groundbreaking work by female scientists, yet the field remains predominantly male, driven by persistent stereotypes about inherent gendered aptitudes..." **Bias Class:** Potential Bias **Bias Category:** Stereotype, Linguistic **Justification:** Emphasises a gender stereotype by attributing women's underrepresentation to innate aptitudes rather than structural barriers, using loaded descriptions of abilities and roles.

		- Retrieved Example 3 (similarity score: 61.8%): **Text Segment:** "In the module on healthcare, a case vignette describes male doctors as 'confident and authoritative' while female nurses are supportive and gentle, reinforcing occupational stereotypes and implying..." **Bias Class:** Bias **Bias Category:** Stereotype **Justification:** Clear occupational stereotypes (stereotyping) link professional authority to men and caring roles to women, reinforcing traditional gender divisions in healthcare.

Table 8. 1:System classification analysis with RAG examples - Example A1

Aspect	Ground-truth Classification	Bias Detection Application Classification
Text Segment	"The development of modern democracy can be traced primarily through European political thought, from ancient Greece to the Enlightenment. Whilst other civilisations contributed to governance, the most significant and influential democratic principles emerged from Western intellectual traditions."	-
Bias Class	Bias	Bias
Bias Category	Representation	Representation Linguistic
Bias Keywords/Phrases	"primarily through European political thought," "most significant and influential democratic principles," "Western intellectual traditions"	"most significant and influential democratic principles emerged from western intellectual traditions", "whilst other civilisations contributed to governance"
Justification	Demonstrates Eurocentric bias and underrepresentation of non-Western contributions to democratic thought.	This text is biased as it perpetuates a Eurocentric view of the development of modern democracy, specifically attributing the most significant and influential democratic principles solely to Western intellectual traditions. The phrase "Whilst other civilisations contributed to governance" is a token acknowledgment, but it is eclipsed by the focus on Western ideas as the primary origin of democracy. The segment overlooks the democratic traditions and contributions from non-Western civilisations, which undermines a comprehensive understanding of the global history of democracy.
Retrieved Examples from RAG Corpus	-	**- Retrieved Example 1 (similarity score: 58.3%):** **Text Segment:** "In the chapter on political systems, one paragraph states that collectivist cultures 'suppress individual creativity in favour of communal harmony,' implicitly praising Western individualism. This fr..." **Bias Class:** Bias **Bias Category:** Representation **Justification:** By portraying collectivist systems as inherently stifling (representational bias) without balance, the text misrepresents alternative political traditions. **- Retrieved Example 2 (similarity score: 58.1%):** **Text Segment:** "Enlightenment philosophers championed secular governance and empirical science in European capitals, framing their ideas as universal truths before exporting them to colonial territories under the gui..." **Bias Class:** Potential Bias **Bias Category:** Representation, Linguistic

		Justification: Centers a Eurocentric narrative that presents Enlightenment thought as inherently universal, overlooking intellectual traditions from other regions and framing export as benevolent. **- Retrieved Example 3 (similarity score: 51.1%):** Text Segment: "Guiding Questions What role does the public play in checking the power of their government? Learning Objectives Evaluate the extent to which an unlimited power of the majority contributes to the heat..." **Bias Class:** Not Bias **Bias Category:** None **Justification:** This segment presents educational objectives and guiding questions designed to encourage critical thinking and analysis of democratic systems, civic institutions, and the role of citizens. The language is neutral, focusing on factual and objective learning outcomes. The content is free from any biased or stereotypical representation, aiming to provide a balanced framework for understanding civic responsibility, governance, and political participation across different levels. There is no indication of bias in the phrasing or the concepts introduced.

Table 8. 2: System classification analysis with RAG examples - Example A2

Aspect	Ground-truth Classification	Bias Detection Application Classification
Text Segment	"During the civil rights movement, activists sometimes resorted to aggressive tactics that disrupted public order. Whilst their cause was understandable, these confrontational approaches often alienated moderate supporters who might have otherwise contributed to peaceful progress."	-
Bias Class	Potential Bias	Potential Bias
Bias Category	Linguistic	Stereotype, Linguistic, Representation
Bias Keywords/Phrases	"resorted to," "aggressive tactics," "disrupted public order," "confrontational approaches," "alienated moderate supporters," "peaceful progress"	"sometimes resorted to aggressive tactics", "these confrontational approaches often alienated moderate supporters"
Justification	Subtle language that frames civil rights activism negatively, using words like "aggressive" and "confrontational."	The language used might imply a negative connotation towards activist approaches during the civil rights movement, potentially biasing the reader's perception of their methods as 'aggressive' and 'disruptive'. This could infringe on an impartial understanding by not equally highlighting the context and reasons for such tactics, possibly leading to a skewed view of the activists and their historical significance.
Retrieved Examples from RAG Corpus	-	**- Retrieved Example 1 (similarity score: 68.6%):** **Text Segment:** "Activity 2. Ordinary People in the Civil Rights Movement Divide students into four groups and assign each group one of the four sections of the We Shall Overcome site; the four sections are entitled ..." **Bias Class:** Not Bias **Bias Category:** Non **Justification:** This segment provides a straightforward instructional activity focused on analysing the contributions of ordinary people to the Civil Rights Movement. It assigns different groups to explore various aspects of the movement (e.g., strategy, cost, prize) without using biased or discriminatory language. The instructions are neutral and inclusive, focusing on understanding the roles of different participants in the movement, and encourage group collaboration. There are no elements of stereotyping, misrepresentation, or biased language in the activity. **- Retrieved Example 2 (similarity score: 68.5%):** **Text Segment:** "Activity 3. Historic Places in the Civil Rights Movement Ask students to form groups again, this time in smaller groups of two or three, and choose a historical site from the We Shall Overcome website..."

		Bias Class: Not Bias **Bias Category:** None **Justification:** This segment presents a neutral and educational activity that encourages students to explore the historical significance of various sites related to the Civil Rights Movement. It promotes research, critical thinking, and an understanding of Martin Luther King Jr.'s influence on the movement, without showing any signs of bias or stereotyping. The language is factual, and the instructions aim to foster an appreciation of both local and national history in an inclusive way. **- Retrieved Example 3 (similarity score: 68%):** **Text Segment:** "Activity 2. Another Perspective on Violence & Civil Rights Begin day two by introducing students to Martin Luther King, Jr.'s stirring rhetoric by having them listen to a brief excerpt from his 'I Hav..." **Bias Class:** Not Bias **Bias Category:** None **Justification:** This segment effectively sets up a structured classroom activity focused on understanding and analysing Martin Luther King, Jr.'s advocacy for nonviolence as a strategy for civil rights, providing a balanced perspective when compared to the previous day's focus on Malcolm X.

Table 8. 3: System classification analysis with RAG examples - Example B1

Aspect	Ground-truth Classification	Bias Detection Application Classification
Text Segment	"Colonial administrators often praised the work ethic of certain immigrant communities, noting how those who successfully assimilated into local customs achieved greater economic prosperity than groups that maintained their traditional practices."	-
Bias Class	Potential Bias	Potential Bias
Bias Category	Stereotype, Linguistic	Stereotype, Linguistic, Representation
Bias Keywords/Phrases	"praised the work ethic," "certain immigrant communities," "successfully assimilated," "greater economic prosperity," "maintained their traditional practices"	"work ethic", "successfully assimilated", "greater economic prosperity", "groups that maintained their traditional practices"
Justification	Implies stereotypes about immigrant success being tied to assimilation and uses evaluative language.	The text segment uses the term "work ethic" which may carry bias as it pertains to the evaluation of one group's cultural practices in comparison to another. It suggests that those who assimilated were more prosperous, which can inherently position assimilation as superior to maintaining traditional practices. The segmentation of groups into those who have assimilated and achieved economic success versus those who have not can present a binary view that overlooks the complexities of cultural identity and the impact of systemic barriers. This language could potentially perpetuate a narrative that values one cultural approach over the other in terms of economic success, which should be critically evaluated in an educational context.
Retrieved Examples from RAG Corpus	-	**- Retrieved Example 1 (similarity score: 58.4%):** **Text Segment**: "Learning objective: By the end of this module, students will be able to critique how 19th-century missionary reports labelled Indigenous cultures as 'primitive' to legitimise colonial expansion, and pr..." **Bias Class:** Potential Bias **Bias Category:** Linguistic, Representation **Justification:** Highlights how a specific term was used to justify bias, encouraging critical analysis of language but itself cites the charged term to demonstrate its historical usage. **- Retrieved Example 2 (similarity score: 53.7%):** **Text Segment:** "Cultural assimilation denotes the process by which immigrant communities adopt the language, norms, and customs of their host nation while encountering systemic barriers such as xenophobic policies a..." **Bias Class:** Not Bias **Bias Category:** None

		Justification: An objective sociocultural definition and analysis presented without evaluative or normative wording. **- Retrieved Example 3 (similarity score: 53.6%):** **Text Segment:** "When discussing colonial history, the module guide describes colonisers as 'civilising agents' who brought progress and modernity to 'backward' regions. It glosses over the violence of conquest and r..." **Bias Class:** Bias **Bias Category:** Representation, Linguistic **Justification:** Referring to colonisers as "civilising agents" and native societies as "backward" (language-tone bias) misrepresents the agency of colonised peoples (representational bias), erasing their complex histories.

Table 8. 4: System classification analysis with RAG examples - Example B2

Aspect	Ground-truth Classification	Bias Detection Application Classification
Text Segment	"The 1920s witnessed's significant changes in women's legal status, including the passage of the 19th Amendment in 1920, which granted women the right to vote. This period also saw increased participation of women in the workforce and higher education institutions."	-
Bias Class	Not Bias	Not Bias
Bias Category	Non	Non
Bias Keywords/Phrases	Non	Non
Justification	Presents factual information without stereotyping, exclusionary language, or biased framing.	The segment is a factual statement regarding historical events with no judgement or bias included.
Retrieved Examples from RAG Corpus	-	**- Retrieved Example 1 (similarity score: 73.4%):** **Text Segment:** "Consider the following discussion questions: What aspects about the home were idealized in late-nineteenth century life? In what ways did women's daily lives begin to change during this period? How d..." **Bias Class:** Not Bias **Bias Category:** Non **Justification:** This segment is an instructional prompt, containing no biased language. Annotators: No bias language is used and people answering these questions will be able to analyse all aspects of the question without being swayed a certain way. No bias in the language used for this passage and its prompts. This is not biased, it asks questions about women's rights in a very generic way for what seems to be educational purposes. **- Retrieved Example 2 (similarity score: 69.4%):** **Text Segment:** "In what activities did the 'New Woman' engage? What new opportunities were available to women in the 1880s-1910s? Working-Class, Immigrant, and African American Women – This part of the activity sho..." **Bias Class:** Not Bias **Bias Category:** None **Justification:** Educational question aimed at exploring historical gender roles. Annotators: This is an educational question aimed at understanding the context of history about New Women. **- Retrieved Example 3 (similarity score: 67.9%):**

		Text Segment: "During the late 19th century, women's grassroots organizations campaigned for expanded educational access, professional careers, and the right to vote, articulating their demands through lectures, pam..." **Bias Class:** Not Bias **Bias Category:** Non **Justification:** A factual summary of social reform efforts, devoid of persuasive or biased terms.

Table 8. 5: System classification analysis with RAG examples - Example C1

Aspect	Ground-truth Classification	Bias Detection Application Classification
Text Segment	"Historians continue to debate the effectiveness of different approaches to social reform. Some scholars argue that gradual institutional change proves more sustainable, whilst others contend that direct action movements create necessary momentum for policy transformation."	-
Bias Class	Not Bias	Not Bias
Bias Category	Non	Non
Bias Keywords/Phrases	Non	Non
Justification	Presents multiple perspectives without favouring one group or using biased language.	Describes a factual debate with no non-factual, biased or stereotyped language.
Retrieved Examples from RAG Corpus	-	**- Retrieved Example 1 (similarity score: 52.2%):** **Text Segment:** "In the chapter on political systems, one paragraph states that collectivist cultures 'suppress individual creativity in favour of communal harmony,' implicitly praising Western individualism. This fr..." **Bias Class:** Bias **Bias Category:** Representation **Justification:** By portraying collectivist systems as inherently stifling (representational bias) without balance, the text misrepresents alternative political traditions. **- Retrieved Example 2 (similarity score: 50.4%):** **Text Segment:** "Integrate evidence from multiple relevant historical sources and interpretations into a reasoned argument about the past. Activity 1. Identifying the Need for Change Begin the lesson by asking student..." **Bias Class:** Not Bias **Bias Category:** None **Justification:** This segment provides a neutral, factual instruction for students to engage with historical documents and speeches by Dr. Martin Luther King, Jr. It does not contain any elements that suggest bias or stereotypes, and it encourages critical thinking through the examination of multiple sources. The language is objective, and there is no indication of any misrepresentation or biased tone. **- Retrieved Example 3 (similarity score: 49.8%):**

		Text Segment: "During the late 19th century, women's grassroots organisations campaigned for expanded educational access, professional careers, and the right to vote, articulating their demands through lectures, pam..." **Bias Class:** Not Bias **Bias Category:** Non **Justification:** A factual summary of social reform efforts, devoid of persuasive or biased terms.

Table 8. 6: System classification analysis with RAG examples - Example C2

Aspect	Ground-truth Classification	Bias Detection Application Classification
Text Segment	"The success of the Underground Railroad demonstrated the courage and resourcefulness of both escaped slaves and their allies. However, historians note that the majority of enslaved people remained in bondage, suggesting that individual acts of resistance, whilst heroic, had limited systemic impact."	-
Bias Class	Potential Bias or Not Bias (debatable)	Not Bias
Bias Category (in case of Potential Bias)	Linguistic	Non
Bias Keywords/Phrases (in case of Potential Bias)	"majority of enslaved people remained in bondage," "individual acts of resistance," "limited systemic impact"	Non
Justification	Could be seen as diminishing the significance of resistance efforts or as providing balanced historical analysis.	The segment provides a balanced view on the impact of the Underground Railroad, acknowledging the bravery of those involved while also highlighting the prevalent oppression faced by the majority of enslaved individuals. The language is neutral and informative, offering a critical historical perspective without veering into bias, stereotyping, or misrepresentation. It is an educational statement that encourages a comprehensive understanding of the era.
Retrieved Examples from RAG Corpus	-	**- Retrieved Example 1 (similarity score: 62%):** **Text Segment:** "Seaman's Protection Certificate for Samuel Fox, August 12, 1854. Black History Collection, Manuscript Division, Library of Congress (2-6) The Underground Railroad Enslaved blacks and their white sym..." **Bias Class:** Not Bias **Bias Category:** Non **Justification:** This segment provides a factual account of the Underground Railroad and the contributions of both enslaved individuals and their allies. The language is neutral and accurate, detailing historical facts about the secret routes and efforts of individuals like William Still. The mention of primary sources, such as maps and narratives, supports the objectivity of the text. There is no stereotyping, misrepresentation, or biased language in this segment, and it is presented in a historically accurate manner. **- Retrieved Example 2 (similarity score: 61.7%):**

		Text Segment: "a narrative recounting my experiences after being captured from my west african homeland, describing the brutal transatlantic voyage and forced labor on a jamaican sugar plantation. i reflect on the a..." **Bias Class:** Not Bias **Bias Category:** Non **Justification:** The passage presents a first-person historical account without evaluative or stereotypical language. **- Retrieved Example 3 (similarity score: 57.3%):** **Text Segment:** "Brown ends his narrative with the story of an attempt to kidnap a fugitive slave family living in Canada that was foiled by a courageous band of African Americans living in Buffalo, who rode to their ..." **Bias Class:** Not Bias **Bias Category:** Non **Justification:** This segment is a neutral presentation of historical content and educational activities. It discusses the narrative of Brown and the significance of a particular event in his story, encouraging students to analyse historical documents related to slavery. There is no language that suggests stereotyping, misrepresentation, or biased tone. The focus is on understanding history through analysis and discussion, which is presented in an unbiased manner.

Table 8. 7: System classification analysis with RAG examples - Example D1

Aspect	Ground-truth Classification	Bias Detection Application Classification
Text Segment	"Traditional family structures in pre-industrial societies typically featured distinct gender roles, with women primarily responsible for domestic duties and child-rearing whilst men engaged in agricultural labour and trade activities."	-
Bias Class	Potential Bias or Not Bias (debatable)	Not Bias
Bias Category (in case of Potential Bias)	Stereotype	Non
Bias Keywords/Phrases (in case of Potential Bias)	"Traditional family structures," "typically featured," "distinct gender roles," "primarily responsible for," "domestic duties," "child-rearing," "agricultural labour," "trade activities"	Non
Justification	Describes historical patterns but could reinforce gender stereotypes depending on context and framing.	Factual statement with a descriptive tone for educational context; no biased language or value judgements applied.
Retrieved Examples from RAG Corpus	-	**- Retrieved Example 1 (similarity score: 65%):** **Text Segment:** "This wise [...] sent me home with solemn advice to 'live as domestic a life as possible,' to 'have but two hours' intelligent life a day,' and 'never to touch pen, brush or pencil again as long as I ..." **Bias Class:** Not Bias **Bias Category:** Non **Justification:** The segment does not provide enough context to determine bias; it seems neutral. **- Retrieved Example 2 (similarity score: 57.6%):** **Text Segment:** "Consider the following discussion questions: What aspects about the home were idealized in late-nineteenth century life? In what ways did women's daily lives begin to change during this period? How d..." **Bias Class:** Not Bias **Bias Category:** None **Justification:** This segment is an instructional prompt, containing no biased language. Annotators: No bias language is used and people answering these questions will be able to analyse all aspects of the question

		without being swayed a certain way. No bias in the language used for this passage and its prompts. This is not biased, it asks questions about women's rights in a very generic way for what seems to be educational purposes. **- Retrieved Example 3 (similarity score: 47.1%):** **Text Segment:** "In what activities did the 'New Woman' engage? What new opportunities were available to women in the 1880s-1910s? Working-Class, Immigrant, and African American Women – This part of the activity sho..." **Bias Class:** Not Bias **Bias Category:** Non **Justification:** Educational question aimed at exploring historical gender roles. Annotators: This is an educational question aimed at understanding the context of history about New Women.

Table 8. 8: System classification analysis with RAG examples - Example D2

8.7 Ethical Considerations

Given the sensitive nature of bias detection in educational content, ethical implications were considered throughout system design and implementation, addressing three key concerns: user data protection, transparency in bias identification, and potential over-reliance on automated decisions in educational contexts.

The system processes user-submitted text excerpts and feedback for classification and system improvement whilst implementing several privacy measures. No Personally Identifiable Information (PII) is collected, all communication between frontend and backend occurs over TLS-encrypted channels, and explicit user consent is obtained via the extension interface. Submitted texts, classification outputs, and feedback are stored securely in AWS DynamoDB with access restricted to authorised research personnel. These practices follow ethical guidelines for educational AI tools and ensure compliance with research data protection norms (Williamson and Eynon, 2020).

The system prioritises interpretability by providing detailed outputs including predicted bias class, associated bias category, justification, contributing bias keywords/phrases and retrieved RAG examples with similarity score. These elements make the model's reasoning visible and accessible, promoting critical engagement rather than passive acceptance. This approach aligns with established principles of explainable AI in contexts where fairness, inclusivity, and interpretation are central (Doshi-Velez and Kim, 2017).

Although the system provides structured and reasoned output, it is not intended to replace human judgement. The detection of bias particularly in complex educational texts remains a subjective and context-dependent task. The system may fail to detect nuanced forms of bias, misinterpret context, or reflect underlying limitations of its training data. To mitigate this risk, the interface includes a built-in user feedback mechanism allowing users to agree or disagree with each result. Additionally, instructional content within the extension explicitly advises users to treat the system's outputs as a form of assistance rather than authority, reinforcing the principle that final interpretation and action remain the responsibility of the human user

8.8 Conclusion

This chapter presented the design, implementation, and evaluation of a web-based bias detection system that operationalised the research methods from previous chapters. The Chrome extension with FastAPI backend successfully integrated the fine-tuned GPT-4o mini model with RAG capabilities to provide accessible bias analysis for educational content.

The pilot study revealed that the system achieved its primary objective of delivering interpretable bias detection. Participants found the interface intuitive and classifications generally understandable, though technical issues with incomplete outputs highlighted areas requiring refinement. The system demonstrated strong alignment with human annotations across clear bias cases and handled ambiguous examples by aligning with valid expert interpretations.

Key limitations included the small pilot sample size, lack of formal accessibility testing, and technical reliability issues. The RAG component enhanced transparency through contextual examples, though varying similarity scores indicated opportunities for corpus improvement.

This work demonstrates that sophisticated AI-driven bias detection can be made accessible to educational practitioners whilst maintaining analytical rigour. The system represents a significant step towards practical implementation of bias detection technologies in higher education, providing a foundation for future research in educational AI applications. However, expanded user testing, technical refinements, and longitudinal impact studies remain necessary for broader deployment.

Chapter 9

9 Conclusion and Future Work

9.1 Conclusion

Gender and ethnic bias in textual data represents a significant challenge within the fields of NLP and AI, as biased language can perpetuate discriminatory attitudes and reinforce societal inequalities. Whilst considerable scholarly attention has been devoted to developing methodologies for identifying and classifying various forms of bias across diverse data types including social media content, news articles and online discourse there remains a notable gap in research specifically addressing the detection of gender and ethnic bias within higher educational texts. This deficiency is particularly concerning given the influential role that academic materials play in shaping students' perceptions and career paths. The identification and mitigation of such biases in educational contexts is therefore crucial, not only for ensuring equitable learning environments but also for addressing the broader implications of discriminatory language in institutional settings. Consequently, the application of NLP and AI techniques to detect gender and ethnic bias in higher education materials represents an essential area for further investigation.

This research addressed the gap by analysing a novel combination of online higher education materials that had not previously been examined together for gender and ethnic bias detection. The study drew on three distinct data sources, thereby providing a comprehensive examination of bias across different dimensions of the digital academic environment. The data sources included: first, a corpus of University of Leeds News Articles (UoL-NA) annotated for gender representation and two subsets annotated for sentiment and subjective bias. Second, three reading list datasets with demographic annotations: two from the University of Leeds first is the Arabic, Islamic, and Middle Eastern Studies Reading List (AIMESRL) and the second is the Disability Studies Reading Lists (DSRL); and one from Imam Mohammad ibn Saud Islamic University, Saudi Arabia the Qur'an Sciences and Islamic Education Reading List (QSIERL). Third, domain-specific higher educational learning resources datasets; first data from Open Educational Resources (OER) Commons focus on learning materials in Humanities disciplines (OER-H) subjects annotated for Representational, Stereotypical and Linguistic bias. Secondly, University of Leeds Data Science and Programming for Data Science lecture transcripts (UoL-DSPDS-LT) used as cross-domain benchmark.

9.1.1 Research Questions Answers

To address the aim and objectives of this research, a main research question was formulated:

How can AI- and NLP-driven frameworks be designed, evaluated and deployed to detect bias across diverse higher education online resources?

Building on this overarching question, a series of supporting research questions were developed to guide the investigation and deconstruct the main inquiry into manageable components:

Q1) What key types of gender and ethnic bias are present in higher education online resources, and how can these be defined and operationalised for detection?

Chapter 3 addressed this question by establishing a comprehensive framework that identified three main bias categories: representation bias, stereotype bias, and linguistic bias. The chapter further detailed how each category manifests within the selected data sources.

- AIMESRL, QSIERL and DSRL: Gender and ethnic – representation bias
- UoL-NA: Gender - stereotype, representation and linguistic
- OER-H: Gender and ethnic - stereotype, representation and linguistic

Furthermore, the chapter provided detailed annotations of each selected dataset for the identified bias categories.

AIMESRL, QSIERL and DSRL

Beginning with manual annotation, author-level gender and ethnicity demographic information was added to the reading lists AIMESRL, QSIERL, and DSRL to investigate representation bias. For each listed author, where possible, efforts were made to identify gender and ethnicity using publicly available academic profiles, institutional biographies, publication records and other online sources. This was further investigated in Chapter 5 through demographic analysis of the reading lists, which revealed that AIMESRL has 82.87% male authorship with Western dominance (44.34% North American and 36.79% European). QSIERL has 95.83% male and 84.72% Middle Eastern authorship. DSRL showed slightly higher female representation (34.5% female, 32.5% male and 0.3% non-binary) though ethnicity for 81.1% was not identifiable.

UoL-NA

In Chapter 3, all three of the identified bias categories linguistic, stereotype and representation bias appear in UoL-NA. A subset of 110 news articles from UoL-NA was annotated for linguistic (subjectivity) bias using MTurk, revealing that out of the random sample, 88 were classified as biased and 22 as not biased. In Chapter 5, representation bias in UoL-NA was investigated by analysing which individuals and groups received coverage in university news. The analysis focused on gender representation patterns, the sentiment expressed in gendered articles. The gender distribution analysis revealed notable disparities in how male and female terms appeared across UoL-

NA. Of the 5,768 articles analysed, 34% predominantly featured male-related terms, 12% featured female-related terms, and the remaining 52% were classified as neutral, containing no dominant gender terms. Chapter 3 described the sentiment annotation process for gendered articles, which Chapter 5 explored further. Sentiment analysis using the zero-shot PLM Facebook BART-Large-MNLI model achieved an overall accuracy of 85%. The model performed well for polarised sentiments (positive F1: 0.89; negative F1: 0.90) but was unable to detect neutral sentiment (F1: 0.00). Two-proportion z-tests found no significant differences in positive ($z = -0.54$, $p = 0.586$) or negative sentiment ($z = 0.48$, $p = 0.632$) between male and female articles. However, this result is limited by the model's failure to identify neutral tones, which accounted for only 0.07% of gendered articles. Stereotype bias is investigated in how each gender was portrayed in relation to career and family themes. Two-proportion z-tests revealed significant differences in the use of career-related terms ($z = 53.81$, $p = 0.000$) and family-related terms ($z = 23.21$, $p = 0.000$) between male and female articles. Proportional analysis showed that female-focused articles contained nearly twice as many family-related terms per article (3.97 vs 2.19), while the frequency of career-related terms remained similar. This indicates a stereotypical tendency to associate women more strongly with family roles in the coverage.

OER-H

Chapter 3 also carried bias annotation for OER-H data for stereotype, representation and linguistic biases. A subset of 140 text segments from OER-H were annotated initially by LLM GPT-4o then was reviewed and re-annotated by human annotators from Prolific. For bias categories GPT-4o annotated 47 instances of stereotype, while the human annotation found 44. For representation, GPT-4o identified 97 cases compared to the human's 83, and for linguistic GPT-4o marked 80 instances, whereas the human annotation recorded 78.

Q2) To what extent can LLMs based bias annotation capture different manifestations of bias across varied higher education online resources?

The two datasets that employed LLMs for bias annotation are UoL-NA and OER-H, both of which were investigated in Chapter 3.

UoL-NA

Within this dataset, GPT-3.5-turbo and Google Bard were evaluated for their ability to detect linguistic (subjectivity) bias. Both models were prompted to classify news articles as either biased or unbiased, and their outputs were compared against human annotations collected via MTurk. The results showed that GPT-3.5-turbo achieved 45% accuracy, while Google Bard achieved 42%, indicating low reliability in identifying biased content. Both models frequently misclassified neutral and subtly biased texts, revealing difficulties in recognising nuanced linguistic bias within university news articles. Overall, these findings suggest that although GPT-3.5-turbo and Google Bard can produce annotations efficiently, their performance in domain-specific bias detection remains limited, lacking the contextual understanding and consistency necessary for dependable analysis.

OER-H

A subset of 140 text segments from the OER-H dataset was initially annotated by LLM GPT-4o and subsequently reviewed and re-annotated by human annotators recruited via Prolific. For comparison purposes, an equal number of GPT-4o annotations was selected from each bias class (Bias, Potential Bias and Not Bias) to ensure a balanced evaluation. However, the human annotators demonstrated a stronger inclination to classify text segments as Not Bias. Of the 140 segments, 59 were classified as 'Not Bias', followed by 45 instances of 'Potential Bias', while 'Bias' was the least assigned category with 36 instances. The comparative analysis revealed strong overall alignment, with an agreement rate of 81.43% between GPT-4o and human annotators. Across individual bias categories, GPT-4o consistently identified slightly more instances than human annotators 47 versus 44 for Stereotype, 97 versus 83 for Representation, and 80 versus 78 for Linguistic bias.

Confidence scores, collected based on the level of agreement among human annotators, provided a quantitative measure of annotation reliability. The analysis of these scores showed that GPT-4o aligned most closely with high-confidence human annotations (≥90%), demonstrating strong performance in detecting explicit forms of bias. However, in lower-confidence cases (≤50%) involving subtle or context-dependent bias, GPT-4o's classifications were less consistent and often lacked sufficient justification. Overall, while GPT-4o exhibited strong concordance with human judgements and performed well in identifying overt bias, it displayed a tendency toward oversensitivity, reinforcing the need for human expertise in interpreting nuanced and contextually embedded bias.

Q3) How effective are zero-shot PLMs, LLMs and fine-tuned LLMs as NLP and AI tools for generating content and detecting bias in higher education online sources?

UoL-NA

In Chapter 4 two Zero-shot PLMs were used as tools to investigate bias in UoL-NA data.

1. The zero-shot PLM DistilBERT achieved an accuracy of 68.8% in detecting subjective (linguistic) bias, demonstrating moderate effectiveness in identifying biased language. It outperformed the zero-shot LLMs GPT-3.5-turbo and Google Bard, which both showed low accuracy below 50%.

2. Sentiment analysis used to reveal representation bias, employing the zero-shot PLM Facebook BART-Large-MNLI, achieved an overall accuracy of 85%, demonstrating strong performance for positive (F1 = 0.89) and negative (F1 = 0.90) sentiment detection. However, the model completely failed to identify neutral sentiment (F1 = 0.00), indicating a significant limitation in distinguishing balanced or non-polarised content. This limitation suggested that although the model performs well in detecting strongly polarised sentiments, its inability to recognise neutral tone highlights a key limitation of the zero-shot approach. Without domain-specific fine-tuning, zero-shot models rely heavily on

generalised linguistic patterns learned from broad datasets, which may not transfer effectively to the more formal and contextually nuanced language typical of university news.

AIMESRL and QSIERL

The LLM GPT-4o was employed in a zero-shot capacity to generate or retrieve standardised summaries of reading list items for subsets within the AIMESRL and QSIERL datasets, ensuring consistent textual representation across all items. The model was prompted to produce concise, academic-style summaries 2–3 sentences, up to 150 tokens with a temperature setting of 0.2 to maintain coherence and minimise variability. A comprehensive validation framework comprising semantic similarity analysis, term preservation analysis, readability assessment and structural evaluation confirmed that the generated summaries demonstrated strong thematic alignment (i.e., mean cosine similarity = 0.743) and adhered closely to established academic writing conventions.

OER-H

In Chapters 6 and 7, LLMs were employed for a range of tasks, including data augmentation, fine-tuning, baseline model evaluation and cross-domain validation.

1. In Chapter 6, LLM GPT-4.o was employed as zero-shot to generate 75 additional educational text examples (25 per bias class) to augment the relatively small OER-H dataset of 108 entries, addressing data scarcity whilst maintaining classification integrity. The model was provided with access to the original annotated data to learn its structure and produce contextually aligned examples reflecting 'Bias', 'Not Bias' and 'Potential Bias' categories, with a maximum of 200 tokens per entry. The generated examples demonstrated strong thematic alignment with the original dataset across key areas including Western intellectual dominance, gender stereotyping, civil rights framing, underrepresentation in literature and immigrant integration, ultimately expanding the dataset to 183 entries whilst preserving the balance and integrity of the human-annotated classifications. Furthermore, in Chapter 7, GPT-4o was applied using the same methodology to perform data augmentation on the RAG corpus.

2. In Chapter 6, following the annotation and data balancing of the OER-H dataset, the LLM GPT-4o mini was fine-tuned on the data and achieved an accuracy of 81.1%, outperforming all baseline zero-shot LLMs. The zero-shot GPT-4o mini attained 67.6% accuracy, while Claude 4 Opus and Claude 4 Sonnet achieved 66.7% and 64.9%, respectively. Cross-domain validation using the UoL-DSPDS-LT dataset revealed minimal bias detection: the fine-tuned GPT-4o mini identified 0.4% of segments as biased, compared to 1.8% identified by the zero-shot GPT-4o mini.

Q4) How can the RAG method be integrated into a bias detection framework for higher education online learning resources, and what advantages does it offer in terms of accuracy, robustness and transparency?

In Chapter 7, the hybrid architecture integrating fine-tuned GPT-4o mini with RAG enhanced transparency and prediction balance by fully eliminating false positives in the 'Not Bias' class, addressing the precision limitation (0.67) observed in the fine-tuned LLM-only approach. The system demonstrated more distributed predictions across bias categories, partially resolving the fine-tuned model's pronounced overprediction of 'Linguistic' bias, though this resulted in a shift towards overpredicting 'Representation' bias rather than eliminating categorical preference entirely. Whilst the hybrid approach achieved a slightly lower overall strict accuracy of 78% compared to the fine-tuned LLM-only model's 81%, it provided more balanced performance across bias classes and categories, offering consistent classification that proved beneficial for detecting diverse bias types in academic texts. The integration of multiple analytical components through RAG enabled the system to emphasise structural and representational elements beyond purely linguistic markers, resulting in a more nuanced, albeit cautious, treatment of borderline cases.

Q5) How can a bias detection framework be implemented in a web application tool to make bias identification methods more accessible and usable within academic environments?

In Chapter 8, the hybrid architecture introduced in Chapter 7 was turned into the practical design of a web application for bias detection in higher education learning materials. The web-based application made bias identification methods more accessible and usable within academic environments by providing an intuitive interface that participants in the study found clear and easy to navigate, with the text input process deemed straightforward and the feedback mechanism very easy to use. The system demonstrated strong operational effectiveness as an educational tool, achieving median response times of 10 seconds for initial submissions and 2 seconds for subsequent analyses, with results consistently aligning with human annotations across diverse bias types. Participants responded positively when asked whether they would consider using the tool in their academic or professional work, indicating 'Yes/Maybe' and suggesting the application demonstrated sufficient utility and reliability for professional contexts. The implementation as a Chrome extension with backend services enabled practical integration into existing academic workflows, allowing users to analyse educational content directly without requiring technical expertise in ML or NLP techniques.

9.1.2 Research Contributions

Addressing the research questions resulted in the following key contributions presented in each chapter:

Chapter 3:

This chapter achieved the main contributions introduced in Chapter 1, namely:

- Established a bias classification framework for online higher education resources (see Section 1.3.1).
- Presented three novel datasets with annotations (see Section 1.3.2).
- Developed LLM-based bias annotation strategies (see Section 1.3.3.1).
- Proposed a hybrid human-LLM annotation methodology (see Section 1.3.4.1).

Chapter 4:

This chapter achieved the main contributions introduced in Chapter 1, namely:

- Initiated the first NLP-based research study to detect gender bias in university news articles (see Section 1.3.4.3).

Chapter 5:

This chapter achieved the main contributions introduced in Chapter 1, namely:

- Designed a replicable NLP pipeline to analyse demographic and thematic diversity in university reading lists (see Section 1.3.4.2).

Chapter 6:

This chapter achieved the main contributions introduced in Chapter 1, namely:

- Fine-tuned a domain-specific bias detection LLM tailored to Humanities disciplines (see Section 1.3.3.2).
- Conducted a cross-domain transfer experiment by applying an LLM trained on humanities data to STEM data, evaluating its capacity to detect bias across distinct disciplinary contexts (see Section 1.3.3.3).

Chapter 7:

This chapter achieved the main contributions introduced in Chapter 1, namely:

- Designed a hybrid framework that integrates a fine-tuned LLM model with RAG for bias detection (see Section 1.3.3.4).

Chapter 8:

This chapter achieved the main contributions introduced in Chapter 1, namely:

- Developed a web application prototype for bias detection within higher education content (see Section 1.3.4.4).

9.1.3 Research Limitations

Although this research has made several significant contributions, it also presents certain limitations that should be acknowledged. This research encountered several methodological, technical and ethical limitations. Chapter 3 identified challenges in demographic annotation of reading lists, where categorising authors into broad ethnic groups risked oversimplifying complex cultural identities and relying on external assumptions that may not have reflected authors lived experiences. The MTurk annotation process demonstrated low inter-annotator agreement, with concerns about worker motivation prioritising speed over accuracy, whilst zero-shot LLM annotations faced variability in predictions and limited domain-specific understanding. GPT-4o demonstrated inconsistency in bias classification, occasionally revising decisions upon re-evaluation, and human annotators showed significant variation in judgements based on personal beliefs, particularly regarding gendered language and controversial historical content.

Chapter 4 identified limitations related to pre-trained NLP models (Facebook BART-Large-MNLI, DistilBERT, Google Bard and GPT-3.5-turbo) that may have reproduced or amplified existing biases without domain-specific fine-tuning. Human annotation proved subject to bias, with low inter-rater agreement and negative Cohen's Kappa scores, whilst the single-institution dataset limited generalisability. Chapter 5 acknowledged that analysis based on summaries rather than full texts may not have captured complete depth and nuance, with LLM-generated summaries introducing abstraction from authentic voices. The small sample size of 40 texts across two departments and moderate clustering metrics (ARI: 0.485, silhouette: 0.137) restricted the breadth and statistical robustness of conclusions.

Chapter 6 revealed that the model overpredicted 'Linguistic' bias with high false positives, frequently confused 'Stereotype' and 'Representation' categories and struggled with multi-label classification. Chapter 7 identified that the RAG corpus was relatively small and heavily influenced by Western educational contexts, potentially missing bias patterns from other cultural settings. The use of strict accuracy metrics, small test sets, and technical limitations including fixed decision thresholds and static prompt construction may not have captured the full complexity of contextual bias detection. Chapter 8 addressed ethical concerns regarding potential over-reliance on automated outputs, noting that bias detection in complex educational texts remained subjective and context-dependent, with the system potentially failing to detect nuanced biases or reflecting training data limitations, necessitating human oversight and critical engagement.

9.2 Future Work

Future research could explore several promising directions. Firstly, expanding the university news dataset to encompass a wider range of institutions both within the UK and internationally would enable the development of a more representative corpus for analysing bias in university-related reporting. Such breadth would also support cross-institutional and cross-cultural comparisons, offering a richer foundation for empirical study.

Secondly, further work could involve fine-tuning a broader selection of LLMs on a more diverse array of higher education learning materials. Expanding beyond the humanities to include engineering, computer science, medicine, law, business and the natural sciences would help assess how domain-specific content influences the emergence or mitigation of bias. This would also clarify whether certain disciplines present unique challenges for bias detection or require bespoke modelling strategies.

Thirdly, the web application could be developed into a more robust platform incorporating wider data sources to support educators in examining potential bias within their teaching materials. A practical extension would be the creation of a browser plug-in such as a Chrome extension capable of integrating with online lecture platforms. This tool could automatically extract and analyse lecture transcripts, slides, or embedded readings in real time, offering immediate feedback on potential sources of bias.

A further avenue concerns deployment in real educational environments. The approach could be implemented as a plug-in or add-on for widely used Learning Management Systems (LMS), such as Blackboard or Minerva, enabling seamless integration into existing teaching workflows. Exploring whether educational agencies, universities, or private companies might adopt the method to produce a deployable product would also be valuable. Such work would necessarily involve addressing issues of data governance, privacy, system interoperability, and long-term maintenance factors that determine whether a research prototype can evolve into a viable industry tool.

Finally, future studies could evaluate the effectiveness of bias-detection systems through structured user studies involving educators, students, learning technologists and policymakers. This would provide insight into usability, interpretability, trust, and the pedagogical implications of surfacing bias within teaching materials. Such evaluations would also help identify failure modes, refine the interface, and guide standards for responsible deployment in educational settings.

References

Abbas, N. and Atwell, E. 2025. Cognitive Computing with Large Language Models for Student Assessment Feedback. *Big Data and Cognitive Computing.* **9**(5), p112.

Abdul-Rahim, R. 2011. *Naskh al-Qur'an: A theological and juridical reconsideration of the theory of abrogation and its impact on Qur'anic exegesis.* Temple University.

Abid, A., Farooqi, M. and Zou, J. 2021. Persistent anti-muslim bias in large language models. In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp.298-306.

AbouElMagd, N. 2016. *Why is my curriculum white? - Decolonising the academy.* [Online]. [Accessed 30 November 2022]. Available from: https://www.nusconnect.org.uk/articles/why-is-my-curriculum-white-decolonising-the-academy

Abram, M.D., Mancini, K.T. and Parker, R.D. 2020. Methods to integrate natural language processing into qualitative research. *International Journal of Qualitative Methods.* **19**, p1609406920984608.

Ahmed, I., Liu, W., Roscoe, R.D., Reilley, E. and McNamara, D.S. 2025. Multifaceted Assessment of Responsible Use and Bias in Language Models for Education. *Computers.* **14**(3), p100.

Ahmed, L. 2021. *Women and gender in Islam: Historical roots of a modern debate.* Yale University Press.

Ahmed, S. 2012. On being included: Racism and diversity in institutional life. *On being included.* Duke University Press.

Aizenberg, E. and Van Den Hoven, J. 2020. Designing for human rights in AI. *Big Data & Society.* **7**(2), p2053951720949566.

al-Nadwi, M.A. 2021. al-Wafa'bi Asma'al-Nisa'Mausu'a Tarajim 'A'lam al-Nisa'fi al-Hadith al-Nabawiyy al-Sharif. *Beirut: Dar al-Najah & Arab Saudi: Dar al-Minhaj.*

Alatas, S.F. 2003. Academic dependency and the global division of labour in the social sciences. *Current sociology.* **51**(6), pp.599-613.

Alayan, S., Rohde, A. and Dhouib, S. 2012. *The politics of education reform in the Middle East: Self and other in textbooks and curricula.* Berghahn Books.

Albuquerque, J. 2023. *Towards an automatic approach for uncovering ethnic bias in online learning texts*. thesis, Open University (United Kingdom).

Albuquerque, J., Rienties, B., Holmes, W. and Hlosta, M. 2025. From hype to evidence: exploring large language models for inter-group bias classification in higher education. *Interactive Learning Environments.* **33**(3), pp.2332-2354.

Alrabaa, S. 1985. Sex division of labour in Syrian school textbooks. *International Review of Education.* **31**(1), pp.335-348.

Alsafari, B., Atwell, E., Walker, A. and Callaghan, M. 2024. Towards effective teaching assistants: From intent-based chatbots to LLM-powered teaching assistants. *Natural Language Processing Journal*. **8**, p100101.

Alwani, Z. 2013. Muslim women as religious scholars: A historical survey. *Muslima theology: The voices of Muslim women theologians*. pp.45-58.

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P.N. and Inkpen, K. 2019. Guidelines for human-AI interaction. In: *Proceedings of the 2019 chi conference on human factors in computing systems*, pp.1-13.

Amutah, C., Greenidge, K., Mante, A., Munyikwa, M., Surya, S.L., Higginbotham, E., Jones, D.S., Lavizzo-Mourey, R., Roberts, D. and Tsai, J. 2021. *Misrepresenting race—the role of medical schools in propagating physician bias*. Mass Medical Soc. 384. pp.872-878.

Apple, M.W. 2014. *Official knowledge: Democratic education in a conservative age*. Routledge.

Aroyo, L. and Welty, C. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*. **36**(1), pp.15-24.

Asr, F.T., Mazraeh, M., Lopes, A., Gautam, V., Gonzales, J., Rao, P. and Taboada, M. 2021. The gender gap tracker: Using natural language processing to measure gender bias in media. *PloS one*. **16**(1), pe0245533.

Bachore, M. and Semela, T. 2022. Is gender-bias in textbooks spilled over from schools to universities? Evidence from Ethiopian university English language course textbooks. *Cogent Education*. **9**(1), p2152616.

Baker, R.S. and Hawn, A. 2021. Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*. pp.1-41.

Banks, J.A. 2015. *Cultural diversity and education: Foundations, curriculum, and teaching*. Routledge.

Bektik, D., Edwards, C., Whitelock, D. and Antonaci, A. 2024. Use of LLM tools within higher education: Report 1.

Beukeboom, C.J. and Burgers, C. 2017. Linguistic bias. *Oxford Encyclopedia of Communication*. Oxford University Press, pp.1-19.

Bhambra, G.K., Gebrial, D. and Nişancıoğlu, K. 2018. *Decolonising the university*. Pluto Press.

Bhat, R.M., Rajan, P. and Gamage, L. 2023. Redressing Historical Bias: Exploring the Path to an Accurate Representation of the Past. *Journal of Social Science*. **4**(3), pp.698-705.

Bhopal, K., Brown, H. and Jackson, J. 2018. Should I stay or should I go? BME academics and the decision to leave UK higher education. *Dismantling race in higher education: Racism, whiteness and decolonising the academy*. Springer, pp.125-139.

Bin Shiha , R., Atwell , E. and Abbas, N. 2022. Decolonising the reading lists of Arabic, Islamic and Middle Eastern Studies. In: *IMAN'2022 International Conference on Islamic Applications in Computer Science And Technology*.

Bin Shiha, R., Atwell, E. and Abbas, N. 2023. Detecting Bias in University News Articles: A Comparative Study Using BERT, GPT-3.5 and Google Bard Annotations. In: *International Conference on Innovative Techniques and Applications of Artificial Intelligence*: Springer, pp.487-492.

Bin Shiha, R., Atwell, E. and Abbas, N. 2024. University News: A New Data Source for NLP Bias Research. In: *International Conference on Innovative Techniques and Applications of Artificial Intelligence*: Springer, pp.307-312.

Blei, D.M., Ng, A.Y. and Jordan, M.I. 2003. Latent dirichlet allocation. *Journal of machine Learning research.* **3**(Jan), pp.993-1022.

Blodgett, S.L., Barocas, S., Daumé Iii, H. and Wallach, H. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050.*

Blumberg, R.L. 2007. Gender bias in textbooks: A hidden obstacle on the road to gender equality in education.

Blumberg, R.L. 2008. The invisible obstacle to educational equality: Gender bias in textbooks. *Prospects.* **38**, pp.345-361.

Bolukbasi, T., Chang, K.-W., Zou, J.Y., Saligrama, V. and Kalai, A.T. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems.* **29**.

Bozkurt, A., Xiao, J., Farrow, R., Bai, J.Y., Nerantzi, C., Moore, S., Dron, J., Stracke, C.M., Singh, L. and Crompton, H. 2024. *The manifesto for teaching and learning in a time of generative AI: A critical collective stance to better navigate the future*. International Council for Open and Distance Education Oslo, Norway. 16. pp.487-513.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G. and Askell, A. 2020. Language models are few-shot learners. *Advances in neural information processing systems.* **33**, pp.1877-1901.

Bucher, M.J.J. and Martini, M. 2024. Fine-tuned'small'LLMs (still) significantly outperform zero-shot generative AI models in text classification. *arXiv preprint arXiv:2406.08660.*

Button, K.S., Ioannidis, J.P., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S. and Munafò, M.R. 2013. Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews neuroscience.* **14**(5), pp.365-376.

Calderon, N., Reichart, R. and Dror, R. 2025. The alternative annotator test for llm-as-a-judge: How to statistically justify replacing human annotators with llms. *arXiv preprint arXiv:2501.10970.*

Campello, R.J., Moulavi, D. and Sander, J. 2013. Density-based clustering based on hierarchical density estimates. In: *Pacific-Asia conference on knowledge discovery and data mining*: Springer, pp.160-172.

Carnes, M., Devine, P.G., Isaac, C., Manwell, L.B., Ford, C.E., Byars-Winston, A., Fine, E. and Sheridan, J. 2012. Promoting institutional change through bias literacy. *Journal of diversity in higher education.* **5**(2), p63.

Cazden, C.B. 1988. *Classroom discourse: The language of teaching and learning.* ERIC.

Chagnon, E., Pandolfi, R., Donatelli, J. and Ushizima, D. 2024. Benchmarking topic models on scientific articles using BERTeley. *Natural Language Processing Journal.* **6**, p100044.

Charles, E. 2019. Decolonizing the curriculum. *Insights.* **32**(1), p24.

Charlesworth, T.E. and Banaji, M.R. 2019. Gender in science, technology, engineering, and mathematics: Issues, causes, solutions. *Journal of Neuroscience.* **39**(37), pp.7228-7243.

Chinta, S.V., Wang, Z., Yin, Z., Hoang, N., Gonzalez, M., Quy, T.L. and Zhang, W. 2024. FairAIED: Navigating fairness, bias, and ethics in educational AI applications. *arXiv preprint arXiv:2407.18745.*

Chisholm, L. 2018. Representations of class, race, and gender in textbooks. *The Palgrave handbook of textbook studies.* pp.225-237.

Christian-Smith, L. and Apple, M. 1991. The politics of the textbook. *New York.*

Cohen, J. 2013. *Statistical power analysis for the behavioral sciences.* routledge.

Cohn, C., Rayala, S., Snyder, C., Fonteles, J., Jain, S., Mohammed, N., Timalsina, U., Burriss, S.K., Srivastava, N. and Deweese, M. 2025. Personalizing Student-Agent Interactions Using Log-Contextualized Retrieval Augmented Generation (RAG). *arXiv preprint arXiv:2505.17238.*

Collins, P.H. 2022. *Black feminist thought: Knowledge, consciousness, and the politics of empowerment.* routledge.

Connell, R. 2020. *Southern theory: The global dynamics of knowledge in social science.* Routledge.

Connor, P., Weeks, M., Glaser, J., Chen, S. and Keltner, D. 2023. Intersectional implicit bias: Evidence for asymmetrically compounding bias and the predominance of target gender. *Journal of Personality and Social Psychology.* **124**(1), p22.

Copur-Gencturk, Y., Thacker, I. and Cimpian, J.R. 2022. Teacher bias in the virtual classroom. *Computers & Education.* **191**, p104627.

Cover, T. and Hart, P. 1967. Nearest neighbor pattern classification. *IEEE transactions on information theory.* **13**(1), pp.21-27.

Crabb, P.B. and Bielawski, D. 1994. The social representation of material culture and gender in children's books. *Sex roles.* **30**, pp.69-79.

Crenshaw, K. 2013. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *Feminist legal theories.* Routledge, pp.23-51.

Crenshaw, K.W. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine (pp. 139–168). In: *University of Chicago legal forum.*

Crowston, K., Allen, E.E. and Heckman, R. 2012. Using natural language processing technology for qualitative data analysis. *International Journal of Social Research Methodology.* **15**(6), pp.523-543.

Cruz, A.F., Rocha, G. and Cardoso, H.L. 2020. On document representations for detection of biased news articles. In: *Proceedings of the 35th annual ACM symposium on applied computing*, pp.892-899.

Csanády, B., Muzsai, L., Vedres, P., Nádasdy, Z. and Lukács, A. 2024. LlamBERT: Large-scale low-cost data annotation in NLP. *arXiv preprint arXiv:2403.15938*.

D'Ignazio, C. and Klein, L. 2020. Data Feminism Cambridge. *MA: MIT Press [Google Scholar]*.

Dacon, J. and Liu, H. 2021. Does gender matter in the news? detecting and examining gender bias in news articles. In: *Companion Proceedings of the Web Conference 2021*, pp.385-392.

Das, A., Zhang, Z., Hasan, N., Sarkar, S., Jamshidi, F., Bhattacharya, T., Rahgouy, M., Raychawdhary, N., Feng, D. and Jain, V. 2024. Investigating annotator bias in large language models for hate speech detection. *arXiv preprint arXiv:2406.11109*.

Davidson, T., Bhattacharya, D. and Weber, I. 2019. Racial bias in hate speech and abusive language detection datasets. *arXiv preprint arXiv:1905.12516*.

de Carvalho, G.H., Knap, O. and Pollice, R. 2024. Show, Don't Tell: Evaluating Large Language Models Beyond Textual Understanding with ChildPlay. *arXiv preprint arXiv:2407.11068*.

Delgado, R. and Stefancic, J. 2023. *Critical race theory: An introduction.* NyU press.

Dev, S., Monajatipoor, M., Ovalle, A., Subramonian, A., Phillips, J.M. and Chang, K.-W. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. *arXiv preprint arXiv:2108.12084*.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp.4171-4186.

Dhamala, J., Sun, T., Kumar, V., Krishna, S., Pruksachatkun, Y., Chang, K.-W. and Gupta, R. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp.862-872.

Divya Venkatesh, J., Jaiswal, A. and Nanda, G. 2024. Comparing human text classification performance and explainability with large language and machine learning models using eye-tracking. *Scientific reports*. **14**(1), p14295.

Dong, X., Wang, Y., Yu, P.S. and Caverlee, J. 2023. Probing explicit and implicit gender bias through llm conditional text generation. *arXiv preprint arXiv:2311.00306*.

Doshi-Velez, F. and Kim, B. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Doughman, J. and Khreich, W. 2022. Gender bias in text: Labeled datasets and lexicons. *arXiv preprint arXiv:2201.08675*.

El-Mesawi, M.E.-T. 2005. The Methodology of al-Tafsīr al-Mawḍū ʿī: A Comparative Analysis. *Intellectual Discourse*. **13**(1).

Elbouanani, A., Dufraisse, E., Tuo, A. and Popescu, A. 2025. CEA-LIST at CheckThat! 2025: evaluating LLMs as detectors of bias and opinion in text. *arXiv preprint arXiv:2507.07539*.

Eldan, R. and Li, Y. 2023. Tinystories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*.

Elmira. 2025. *Flesch Reading Ease Formula: A Complete Guide.* [Online]. [Accessed 4 September 2025]. Available from: https://clickhelp.com/clickhelp-technical-writing-blog/flesch-reading-ease-formula-a-complete-guide/

Emirtekin, E. 2025. Large language model-powered automated assessment: a systematic review. *Applied Sciences.* **15**(10), p5683.

Equality and Human Rights Commission. 2021. *Protected characteristics, Equality Act 2010.* [Online]. [Accessed 24 November 2025]. Available from: https://www.equalityhumanrights.com/equality/equality-act-2010/protected-characteristics

Fan, Y., Shepherd, L.J., Slavich, E., Waters, D., Stone, M., Abel, R. and Johnston, E.L. 2019. Gender and cultural bias in student evaluations: Why representation matters. *PloS one.* **14**(2), pe0209749.

Fan, Z., Chen, R., Xu, R. and Liu, Z. 2024. Biasalert: A plug-and-play tool for social bias detection in llms. *arXiv preprint arXiv:2407.10241*.

Fasching, N. and Lelkes, Y. 2025. Model-Dependent Moderation: Inconsistencies in Hate Speech Detection Across LLM-based Systems. In: *Findings of the Association for Computational Linguistics: ACL 2025*, pp.22271-22285.

Field, A. 2024. *Discovering statistics using IBM SPSS statistics.* Sage publications limited.

Field, A., Blodgett, S.L., Waseem, Z. and Tsvetkov, Y. 2021. A survey of race, racism, and anti-racism in NLP. *arXiv preprint arXiv:2106.11410*.

Flesch, R. 1948. A new readability yardstick. *Journal of applied psychology.* **32**(3), p221.

Fort, K., Adda, G. and Cohen, K.B. 2011. Amazon Mechanical Turk: Gold mine or coal mine? *Computational Linguistics.* **37**(2), pp.413-420.

Foster, S.J. 1999. The struggle for American identity: Treatment of ethnic groups in United States history textbooks. *History of Education.* **28**(3), pp.251-278.

Gallegos, I.O., Rossi, R.A., Barrow, J., Tanjim, M.M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R. and Ahmed, N.K. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics.* **50**(3), pp.1097-1179.

Garg, N., Schiebinger, L., Jurafsky, D. and Zou, J. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences.* **115**(16), pp.E3635-E3644.

Garrido-Muñoz, I., Montejo-Ráez, A., Martínez-Santiago, F. and Ureña-López, L.A. 2021. A survey on bias in deep NLP. *Applied Sciences.* **11**(7), p3184.

Gaucher, D., Friesen, J. and Kay, A.C. 2011. Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of personality and social psychology.* **101**(1), p109.

Gautam, S. and Srinath, M. 2024. Blind spots and biases: Exploring the role of annotator cognitive biases in nlp. *arXiv preprint arXiv:2404.19071*.

George, M. 1981. *The rise of colleges: institutions of learning in Islam and the West*. Edinburg Universitiy Press.

Gezici, G. and Saygin, Y. 2022. Measuring gender bias in educational videos: A case study on youtube. *arXiv preprint arXiv:2206.09987*.

Gilardi, F., Alizadeh, M. and Kubli, M. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*.

Giorgi, T., Cima, L., Fagni, T., Avvenuti, M. and Cresci, S. 2025. Human and LLM biases in hate speech annotations: A socio-demographic analysis of annotators and targets. In: *Proceedings of the International AAAI Conference on Web and Social Media*, pp.653-670.

Gooden, A.M. and Gooden, M.A. 2001. Gender representation in notable children's picture books: 1995–1999. *Sex roles*. **45**, pp.89-101.

Görke, A. and Pink, J. 2014. Tafsır and Islamic Intellectual History.

Grootendorst, M. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Guo, Y., Guo, M., Su, J., Yang, Z., Zhu, M., Li, H., Qiu, M. and Liu, S.S. 2024. Bias in large language models: Origin, evaluation, and mitigation. *arXiv preprint arXiv:2411.10915*.

Gupta, A.F. and Yin, A.L.S. 1990. Gender representation in English language textbooks used in the Singapore primary schools. *Language and education*. **4**(1), pp.29-50.

Gyamera, G.O. and Burke, P.J. 2018. Neoliberalism and curriculum in higher education: A post-colonial analyses. *Teaching in Higher Education*. **23**(4), pp.450-467.

Hahsler, M., Piekenbrock, M. and Doran, D. 2019. dbscan: Fast density-based clustering with R. *Journal of Statistical Software*. **91**, pp.1-30.

Hall, J., Velickovic, V. and Rajapillai, V. 2021. Students as partners in decolonising the curriculum. *The Journal of Educational Innovation, Partnership and Change*. **7**(1).

Haq, M.U.U., Rigoni, D. and Sperduti, A. 2025. Llms as data annotators: How close are we to human performance. *arXiv preprint arXiv:2504.15022*.

Harding, S. 1991. *Whose science? Whose knowledge?: Thinking from women's lives*. Cornell University Press.

Hartley, J. 2003. Improving the clarity of journal abstracts in psychology: the case for structure. *Science Communication*. **24**(3), pp.366-379.

Hartley, M. and Morphew, C.C. 2008. What's being sold and to what end? A content analysis of college viewbooks. *The Journal of Higher Education*. **79**(6), pp.671-691.

He, X., Lin, Z., Gong, Y., Jin, A., Zhang, H., Lin, C., Jiao, J., Yiu, S.M., Duan, N. and Chen, W. 2023. Annollm: Making large language models to be better crowdsourced annotators. *arXiv preprint arXiv:2303.16854*.

Hinton, P. 2017. Implicit stereotypes and the predictive brain: cognition and culture in "biased" person perception. *Palgrave Communications*. **3**(1), pp.1-9.

Hitti, Y., Jang, E., Moreno, I. and Pelletier, C. 2019. Proposed taxonomy for gender bias in text; a filtering methodology for the gender generalization subtype. In: *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pp.8-17.

Hlatshwayo, M.N. and Fomunyam, K.G. 2019. Theorising the# MustFall student movements in contemporary South African higher education: A social justice perspective. *Journal of Student Affairs in Africa*. **7**(1), pp.61-80.

Holmes, W., Bialik, M. and Fadel, C. 2019. *Artificial intelligence in education promises and implications for teaching and learning*. Center for Curriculum Redesign.

Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M. and Wallach, H. 2019. Improving fairness in machine learning systems: What do industry practitioners need? In: *Proceedings of the 2019 CHI conference on human factors in computing systems*, pp.1-16.

Hovy, D. and Prabhumoye, S. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*. **15**(8), pe12432.

Howard, H.A., Bochenek, A., Mayhook, Z., Trowbridge, T. and Lux, S. 2023. Student information use during the COVID-19 pandemic. *The Journal of Academic Librarianship*. **49**(3), p102696.

Hsu, F.-h. 1992. *A case study of discriminatory gender stereotyping in Taiwan's elementary school textbooks*. Teachers College, Columbia University.

Hsueh, P.-Y., Melville, P. and Sindhwani, V. 2009. Data quality from crowdsourcing: a study of annotation selection criteria. In: *Proceedings of the NAACL HLT 2009 workshop on active learning for natural language processing*, pp.27-35.

Hu, T., Kyrychenko, Y., Rathje, S., Collier, N., van der Linden, S. and Roozenbeek, J. 2025. Generative language models exhibit social identity biases. *Nature Computational Science*. **5**(1), pp.65-75.

Hu, X. and Hancock, A. 2024. State of the science: Implicit Bias in Education 2018–2020. *The Kirwan Institute for the Study of Race and Ethnicity*.

Huang, F., Kwak, H. and An, J. 2023. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. *arXiv preprint arXiv:2302.07736*.

Huang, P. and Liu, X. 2024. Challenging gender stereotypes: representations of gender through social interactions in English learning textbooks. *Humanities and Social Sciences Communications*. **11**(1), pp.1-14.

Huang, T.-H., Cao, C., Bhargava, V. and Sala, F. 2024. The alchemist: Automated labeling 500x cheaper than llm data annotators. *Advances in Neural Information Processing Systems*. **37**, pp.62648-62672.

Huang, Y. 2024. Unveiling Gender Bias in Large Language Models: Using Teacher's Evaluation in Higher Education As an Example. *arXiv preprint arXiv:2409.09652*.

Hube, C. and Fetahu, B. 2018. Detecting biased statements in wikipedia. In: *Companion proceedings of the the web conference 2018*, pp.1779-1786.

Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y. and Denuyl, S. 2020. Social biases in NLP models as barriers for persons with disabilities. *arXiv preprint arXiv:2005.00813*.

Ijoma, J.N., Sahn, M., Mack, K.N., Akam, E., Edwards, K.J., Wang, X., Surpur, A. and Henry, K.E. 2022. Visions by WIMIN: BIPOC representation matters. *Molecular imaging and biology*. **24**(3), pp.353-358.

Islam, K.M.M. and Asadullah, M.N. 2018. Gender stereotypes and education: A comparative content analysis of Malaysian, Indonesian, Pakistani and Bangladeshi school textbooks. *PloS one*. **13**(1), pe0190807.

Iyyer, M., Enns, P., Boyd-Graber, J. and Resnik, P. 2014. Political ideology detection using recursive neural networks. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp.1113-1122.

Jackie, F. and Lee, K. 2011. *Gender Representation--An Exploration of Standardized Evaluation Methods*. Springer Nature BV.

Jehle, A.M., Groeneveld, M.G., van de Rozenberg, T.M. and Mesman, J. 2024. The hidden lessons in textbooks: Gender representation and stereotypes in European mathematics and language books. *European Journal of Education*. **59**(4), pe12716.

Jha, J., Ghatak, N., Menon, N., Dutta, P. and Mahendiran, S. 2019. *Women's education and empowerment in rural India*. Routledge.

Jin, Y. 2002. A discussion on the reform of elementary school social teaching materials from the angle of gender analysis. *Chinese Education & Society*. **35**(5), pp.63-76.

Johnson, J.M. and Khoshgoftaar, T.M. 2019. Survey on deep learning with class imbalance. *Journal of big data*. **6**(1), pp.1-54.

Kasy, M. and Abebe, R. 2021. Fairness, equality, and power in algorithmic decision-making. In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp.576-586.

Kausar, G., Saleem, S., Subhan, F., Suud, M.M., Alam, M. and Uddin, M.I. 2023. Prediction of gender-biased perceptions of learners and teachers using machine learning. *Sustainability*. **15**(7), p6241.

Keith, M.G., Tay, L. and Harms, P.D. 2017. Systems perspective of Amazon Mechanical Turk for organizational research: Review and recommendations. *Frontiers in psychology*. **8**, p1359.

Khokhlova, O., Lamba, N. and Kishore, S. 2023. Evaluating student evaluations: Evidence of gender bias against women in higher education based on perceived learning and instructor personality. In: *Frontiers in education*: Frontiers Media SA, p.1158132.

Kim, D. 2013. *Categorical data analysis, by AGRESTI, ALAN*. Oxford University Press.

Kim, H., Mitra, K., Chen, R.L., Rahman, S. and Zhang, D. 2024. Meganno+: A human-llm collaborative annotation system. *arXiv preprint arXiv:2402.18050*.

Kojima, T., Gu, S.S., Reid, M., Matsuo, Y. and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*. **35**, pp.22199-22213.

Kropman, M., van Drie, J. and Van Boxtel, C. 2023. The influence of multiperspectivity in history texts on students' representations of a historical event. *European Journal of Psychology of Education*. **38**(3), pp.1295-1315.

Kumar, G. and Singh, J.P. 2024. Bias mitigation in text classification through cGAN and LLMs. *Indian National Science Academy Proceedings*.

Kuratomi, G., Pirozelli, P., Cozman, F.G. and Peres, S.M. 2025. A RAG-Based Institutional Assistant. *arXiv preprint arXiv:2501.13880*.

Kurita, K., Vyas, N., Pareek, A., Black, A.W. and Tsvetkov, Y. 2019. Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337*.

Kuzman, T., Ljubešić, N. and Mozetič, I. 2023a. Chatgpt: Beginning of an end of manual annotation? use case of automatic genre identification. *arXiv preprint arXiv:2303.03953*.

Kuzman, T., Mozetič, I. and Ljubešić, N. 2023b. Chatgpt: Beginning of an end of manual linguistic data annotation? use case of automatic genre identification. *arXiv preprint arXiv:2303.03953*.

Laakso, L. and Hallberg Adu, K. 2024. 'The unofficial curriculum is where the real teaching takes place': faculty experiences of decolonising the curriculum in Africa. *Higher Education*. **87**(1), pp.185-200.

Lange, R.A. and Kelley, W.T. 1971. The problem of bias in the writing of elementary history textbooks. *The Journal of General Education*. pp.257-267.

Larson, J., Mattu, S., Kirchner, L. and Angwin, J. 2017. *How We Analyzed the COMPAS Recidivism Algorithm. ProPublica 23 May 2016*.

Lau, J.H., Newman, D. and Baldwin, T. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp.530-539.

Ldli, G. and Zhenzhou, Z. 2002. Children, gender, and language teaching materials. *Chinese Education & Society*. **35**(5), pp.34-52.

le Grange, L. 2016. *â€˜ Decolonising the university curriculum. â€™ South African Journal of Higher Education, 30 (2), 1-12*.

Lee, J., Hicke, Y., Yu, R., Brooks, C. and Kizilcec, R.F. 2024. The life cycle of large language models in education: A framework for understanding sources of bias. *British Journal of Educational Technology*. **55**(5), pp.1982-2002.

Lee, J.F. and Collins, P. 2008. Gender voices in Hong Kong English textbooks—Some past and current practices. *Sex Roles*. **59**, pp.127-137.

Leeson, W., Resnick, A., Alexander, D. and Rovers, J. 2019. Natural language processing (NLP) in qualitative public health research: a proof of concept study. *International Journal of Qualitative Methods*. **18**, p1609406919887021.

Leslie, S.-J., Cimpian, A., Meyer, M. and Freeland, E. 2015. Expectations of brilliance underlie gender distributions across academic disciplines. *Science*. **347**(6219), pp.262-265.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V. and Zettlemoyer, L. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t. and Rocktäschel, T. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*. **33**, pp.9459-9474.

Li, Z., Wang, Z., Wang, W., Hung, K., Xie, H. and Wang, F.L. 2025. Retrieval-augmented generation for educational application: A systematic survey. *Computers and Education: Artificial Intelligence*. p100417.

Liang, P.P., Wu, C., Morency, L.-P. and Salakhutdinov, R. 2021. Towards understanding and mitigating social biases in language models. In: *International conference on machine learning*: PMLR, pp.6565-6576.

Lili, Z. 2003. A Study of Junior Middle School Language Teaching Gender Issues Background to the Study. *Chinese Education & Society*. **36**(3), pp.19-26.

Lin, L., Wang, L., Guo, J. and Wong, K.-F. 2024. Investigating bias in llm-based bias detection: Disparities between llms and human perception. *arXiv preprint arXiv:2403.14896*.

Liu, C., Hoang, L., Stolman, A., Kizilcec, R.F. and Wu, B. 2025. Investigating Student Interaction Patterns with Large Language Model-Powered Course Assistants in Computer Science Courses. *arXiv preprint arXiv:2509.08862*.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R. and Zhu, C. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Llorens, A., Tzovara, A., Bellier, L., Bhaya-Grossman, I., Bidet-Caulet, A., Chang, W.K., Cross, Z.R., Dominguez-Faus, R., Flinker, A. and Fonken, Y. 2021. Gender bias in academia: A lifetime problem that needs solutions. *Neuron*. **109**(13), pp.2047-2074.

Loewen, J.W. 2007. *Lies my teacher told me: Everything your American history textbook got wrong*. Simon and Schuster.

Lu, H., Isonuma, M., Mori, J. and Sakata, I. 2024. Towards Transfer Unlearning: Empirical Evidence of Cross-Domain Bias Mitigation. *arXiv preprint arXiv:2407.16951*.

Lucy, L., Demszky, D., Bromley, P. and Jurafsky, D. 2020. Content analysis of textbooks via natural language processing: Findings on gender, race, and ethnicity in Texas US history textbooks. *AERA Open*. **6**(3), p2332858420940312.

Maass, A. 1999. Linguistic intergroup bias: Stereotype perpetuation through language. *Advances in experimental social psychology*. Elsevier, pp.79-121.

Mahmud, A. and Gagnon, J. 2023. Racial disparities in student outcomes in British higher education: Examining mindsets and bias. *Teaching in Higher Education.* **28**(2), pp.254-269.

Masoudian, S., Escobedo, G., Strauss, H. and Schedl, M. 2025. Investigating Gender Bias in LLM-Generated Stories via Psychological Stereotypes. *arXiv preprint arXiv:2508.03292*.

Matfield, K. 2016. *Gender Decoder: find subtle bias in job ads.* [Online]. [Accessed 09-06-2023]. Available from: http://gender-decoder.katmatfield.com/

McAuliffe, J.D. 1991. *Qur'ānic Christians: An analysis of classical and modern exegesis.* Cambridge University Press.

McCullagh, C.B. 2000. Bias in historical description, interpretation, and explanation. *History and Theory.* **39**(1), pp.39-66.

McHugh, M.L. 2013. The chi-square test of independence. *Biochemia medica.* **23**(2), pp.143-149.

McInnes, L., Healy, J. and Astels, S. 2017. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.* **2**(11), p205.

McInnes, L., Healy, J. and Melville, J. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Měchura, M. 2022. A taxonomy of bias-causing ambiguities in machine translation. In: *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pp.168-173.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR).* **54**(6), pp.1-35.

Mernissi, F. 1991. *The veil and the male elite: A feminist interpretation of women's rights in Islam.* Perseus Books Cambridge, MA.

Intersectionality. 2021. s.v.

Mikołajczyk-Bareła, A. and Grochowski, M. 2025. A survey on bias in machine learning research. *IEEE Access.* **14**, pp.3284-3311.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mimno, D., Wallach, H., Talley, E., Leenders, M. and McCallum, A. 2011. Optimizing semantic coherence in topic models. In: *Proceedings of the 2011 conference on empirical methods in natural language processing*, pp.262-272.

Mlama, P. 2005. Pressure from Within: The Forum for African Women Educationalists. *Partnerships for girls' education.* **49**.

Moore, M.R. 2017. Women of color in the academy: Navigating multiple intersections and multiple hierarchies. *Social Problems.* **64**(2), pp.200-205.

Moss-Racusin, C.A., Dovidio, J.F., Brescoll, V.L., Graham, M.J. and Handelsman, J. 2012. Science faculty's subtle gender biases favor male students. *Proceedings of the national academy of sciences.* **109**(41), pp.16474-16479.

Munoz-Najar, A., Gilberto, A., Hasan, A., Cobo, C., Azevedo, J.P. and Akmal, M. 2021. Remote Learning during COVID-19: Lessons from Today, Principles for Tomorrow. *World Bank*.

Nangia, N., Vania, C., Bhalerao, R. and Bowman, S.R. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133*.

National Union of Students 2019. *"Why is my curriculum white?"*. [Online]. [Accessed 30 November 2022]. Available from: http://www.dtmh.ucl.ac.uk/videos/curriculum-white/

Nemkova, P.A., Ubani, S. and Albert, M.V. 2025. Comparing LLM Text Annotation Skills: A Study on Human Rights Violations in Social Media Data. *arXiv preprint arXiv:2505.10260*.

Newman, D., Lau, J.H., Grieser, K. and Baldwin, T. 2010. Automatic evaluation of topic coherence. In: *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, pp.100-108.

OHCHR. no date. *Gender stereotyping*. [Online]. [Accessed 12 September 2025]. Available from: https://www.ohchr.org/en/women/gender-stereotyping

Ong, M., Wright, C., Espinosa, L. and Orfield, G. 2011. Inside the double bind: A synthesis of empirical research on undergraduate and graduate women of color in science, technology, engineering, and mathematics. *Harvard educational review*. **81**(2), pp.172-209.

Opitz, J. 2022. From Bias and Prevalence to Macro F1, Kappa, and MCC: A structured overview of metrics for multi-class evaluation. *Heidelberg University*.

Opitz, J. 2024. A closer look at classification evaluation metrics and a critical reflection of common evaluation practice. *transactions of the association for computational linguistics*. **12**, pp.820-836.

Orgeira-Crespo, P., Míguez-Álvarez, C., Cuevas-Alonso, M. and Rivo-López, E. 2021. An analysis of unconscious gender bias in academic texts by means of a decision algorithm. *Plos one*. **16**(9), pe0257903.

Institutional bias. 2025. s.v.

Palan, S. and Schitter, C. 2018. Prolific. ac—A subject pool for online experiments. *Journal of behavioral and experimental finance*. **17**, pp.22-27.

Pangakis, N. and Wolken, S. 2024a. Keeping humans in the loop: Human-centered automated annotation with generative ai. *arXiv preprint arXiv:2409.09467*.

Pangakis, N. and Wolken, S. 2024b. Knowledge distillation in automated annotation: Supervised text classification with LLM-generated training labels. *arXiv preprint arXiv:2406.17633*.

Pannucci, C.J. and Wilkins, E.G. 2010. Identifying and avoiding bias in research. *Plastic and reconstructive surgery*. **126**(2), pp.619-625.

Park, J., Wisniewski, P. and Singh, V. 2024. Leveraging large language models (llms) to support collaborative human-ai online risk data annotation. *arXiv preprint arXiv:2404.07926*.

Park, J.K., Ellezhuthil, R.D., Wisniewski, P. and Singh, V. 2024. Collaborative human-AI risk annotation: co-annotating online incivility with CHAIRA. *arXiv preprint arXiv:2409.14223*.

Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., Htut, P.M. and Bowman, S.R. 2021. BBQ: A hand-built bias benchmark for question answering. *arXiv preprint arXiv:2110.08193*.

Pavlick, E. and Kwiatkowski, T. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*. **7**, pp.677-694.

Pavlovic, M. and Poesio, M. 2024. The effectiveness of LLMs as annotators: A comparative overview and empirical analysis of direct representation. *arXiv preprint arXiv:2405.01299*.

Peterson, D.A., Biederman, L.A., Andersen, D., Ditonto, T.M. and Roe, K. 2019. Mitigating gender bias in student evaluations of teaching. *PloS one*. **14**(5), pe0216241.

Phull, K., Ciflikli, G. and Meibauer, G. 2019. Gender and bias in the International Relations curriculum: Insights from reading lists. *European Journal of International Relations*. **25**(2), pp.383-407.

Ping, C. and Weiling, C. 2002. Separated by one layer of paper: Some views about the inferior position of girls in the study of elementary school arithmetic. *Chinese Education & Society*. **35**(5), pp.23-33.

Plank, B. 2022. The'Problem'of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. *arXiv preprint arXiv:2211.02570*.

Poesio, M. and Artstein, R. 2005. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In: *Proceedings of the workshop on frontiers in corpus annotations ii: Pie in the sky*, pp.76-83.

Prates, M.O., Avelar, P.H. and Lamb, L.C. 2020. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*. **32**(10), pp.6363-6381.

Purdie-Vaughns, V. and Eibach, R.P. 2008. Intersectional invisibility: The distinctive advantages and disadvantages of multiple subordinate-group identities. *Sex roles*. **59**(5), pp.377-391.

Purdue University OWL no date. *Stereotypes and Biased Language*. [Online]. [Accessed 12 September 2025]. Available from: https://owl.purdue.edu/owl/general_writing/academic_writing/using_appropriate_language/stereotypes_and_biased_language.html

Pustejovsky, J. and Stubbs, A. 2012. *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. " O'Reilly Media, Inc.".

Quinn, D.M. 2020. Experimental evidence on teachers' racial bias in student evaluation: The role of grading scales. *Educational Evaluation and Policy Analysis*. **42**(3), pp.375-392.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I. 2019. Language models are unsupervised multitask learners. *OpenAI blog*. **1**(8), p9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. and Liu, P.J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*. **21**(140), pp.1-67.

Ramírez, S. 2018. FastAPI: The modern, fast (high-performance) web framework for building APIs with Python 3.7+. *FastAPI Documentation*.

Ranjan, R., Gupta, S. and Singh, S.N. 2024. A comprehensive survey of bias in llms: Current landscape and future directions. *arXiv preprint arXiv:2409.16430*.

Rathbun, G. and Turner, N. 2012. Authenticity in academic development: the myth of neutrality. *International Journal for Academic Development*. **17**(3), pp.231-242.

Raza, S., Reji, D.J., Liu, D.D., Bashir, S.R. and Naseem, U. 2022. An Approach to Ensure Fairness in News Articles. *arXiv preprint arXiv:2207.03938*.

Read, B., Francis, B. and Robson, J. 2005. Gender,'bias', assessment and feedback: Analyzing the written assessment of undergraduate history essays. *Assessment & Evaluation in Higher Education*. **30**(3), pp.241-260.

Reimers, N. and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Rekabsaz, N. and Schedl, M. 2020. Do neural ranking models intensify gender bias? In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.2065-2068.

Ren, R., Ma, J. and Zheng, Z. 2025. Large language model for interpreting research policy using adaptive two-stage retrieval augmented fine-tuning method. *Expert Systems with Applications*. **278**, p127330.

Rist, T. 2023. Confirmation bias studies: towards a scientific theory in the humanities. *SN Social Sciences*. **3**(8), p123.

Röder, M., Both, A. and Hinneburg, A. 2015. Exploring the space of topic coherence measures. In: *Proceedings of the eighth ACM international conference on Web search and data mining*, pp.399-408.

Rodriguez, J. 2018. Why is my curriculum white? In: *10th International Critical Management Studies Conference: Time for another revolution?*

Romanowski, M. 1996. Problems of bias in history textbooks. *Social Education*. **60**, pp.170-173.

Rouse, S.V. 2019. Reliability of MTurk data from masters and workers. *Journal of Individual Differences*.

Rouzegar, H. and Makrehchi, M. 2024. Enhancing text classification through llm-driven active learning and human annotation. *arXiv preprint arXiv:2406.12114*.

Rudinger, R., Naradowsky, J., Leonard, B. and Van Durme, B. 2018. Gender bias in coreference resolution. *arXiv preprint arXiv:1804.09301*.

Sadker, D. and Zittleman, K. 2007. Gender bias: From colonial America to today's classrooms. *Multicultural education: Issues and perspectives*. pp.135-169.

Sadker, M.P. and Sadker, D.M. 1980. Sexism in teacher-education texts. *Harvard Educational Review*. **50**(1), pp.36-46.

Said, E.W. 1978. Orientalism. New York: Pantheon Books.

Saliba, G. 2007. *Islamic science and the making of the European renaissance*. Mit Press.

Samsir, S., Saragih, R.S., Subagio, S., Aditiya, R. and Watrianthos, R. 2023. BERTopic modeling of natural language processing abstracts: Thematic structure and trajectory. *Jurnal Media Informatika Budidarma.* **7**(3), pp.1514-1520.

Sanh, V., Debut, L., Chaumond, J. and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Santos, J., Bittencourt, I., Reis, M., Chalco, G. and Isotani, S. 2022. Two billion registered students affected by stereotyped educational environments: an analysis of gender-based color bias. *Humanities and Social Sciences Communications.* **9**(1), pp.1-16.

Sap, M., Card, D., Gabriel, S., Choi, Y. and Smith, N.A. 2019. The risk of racial bias in hate speech detection. In: *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp.1668-1678.

Sap, M., Swayamdipta, S., Vianna, L., Zhou, X., Choi, Y. and Smith, N.A. 2021. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *arXiv preprint arXiv:2111.07997*.

Saunders, D. and Byrne, B. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. *arXiv preprint arXiv:2004.04498*.

Schoeman, S. 2009. The representation of women in a sample of post-1994 South African school History textbooks. *South African Journal of Education.* **29**(4), pp.541-556.

Schucan Bird, K. and Pitman, L. 2020. How diverse is your reading list? Exploring issues of representation and decolonisation in the UK. *Higher Education.* **79**(5), pp.903-920.

Sesko, A.K. and Biernat, M. 2010. Prototypes of race and gender: The invisibility of Black women. *Journal of Experimental Social Psychology.* **46**(2), pp.356-360.

Shah, D., Schwartz, H.A. and Hovy, D. 2019. Predictive biases in natural language processing models: A conceptual framework and overview. *arXiv preprint arXiv:1912.11078*.

Shah, D.J., Lei, T., Moschitti, A., Romeo, S. and Nakov, P. 2018. Adversarial domain adaptation for duplicate question detection. *arXiv preprint arXiv:1809.02255*.

Shahbazi, N., Lin, Y., Asudeh, A. and Jagadish, H. 2023. Representation bias in data: A survey on identification and resolution techniques. *ACM Computing Surveys.* **55**(13s), pp.1-39.

Shannon, C.E. 1948. A mathematical theory of communication. *The Bell system technical journal.* **27**(3), pp.379-423.

Sheltzer, J.M. and Smith, J.C. 2014. Elite male faculty in the life sciences employ fewer women. *Proceedings of the National Academy of Sciences.* **111**(28), pp.10107-10112.

Sheng, E., Chang, K.-W., Natarajan, P. and Peng, N. 2019. The woman worked as a babysitter: On biases in language generation. *arXiv preprint arXiv:1909.01326*.

Shiha, R.B., Atwell, E. and Abbas, N. 2025a. Bias Detection in Online Higher Education Texts: Comparing Fine-Tuned and Zero-Shot LLM Approaches. In: *International Conference on Innovative Techniques and Applications of Artificial Intelligence*: Springer, pp.285-290.

Shiha, R.B., Atwell, E. and Abbas, N. 2025b. A Transformer-Based Framework for Thematic Analysis of University Reading Lists. In: *International Conference on Innovative Techniques and Applications of Artificial Intelligence*: Springer, pp.405-410.

Shin, N., Forrest, E., Sauceda, A. and Harvey, D. no date. Grappling with Linguistic Bias in the Classroom.

Sleeter, C.E. and Grant, C.A. 2008. *Making choices for multicultural education: Five approaches to race, class and gender.* John Wiley & Sons.

Smith, E.M., Hall, M., Kambadur, M., Presani, E. and Williams, A. 2022. " I'm sorry to hear that": Finding New Biases in Language Models with a Holistic Descriptor Dataset. *arXiv preprint arXiv:2205.09209*.

Smith, J., Meyer, F. and McClure, H. 2023. The persistence of bias in education: A call for research to move policy and practice from aspiration to results. *Policy Futures in Education.* p14782103231180423.

Song, K., Tan, X., Qin, T., Lu, J. and Liu, T.-Y. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems.* **33**, pp.16857-16867.

Soundararajan, S. and Delany, S.J. 2024. Investigating gender bias in large language models through text generation. In: *Proceedings of the 7th international conference on natural language and speech processing (icnlsp 2024)*, pp.410-424.

Sparck Jones, K. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation.* **28**(1), pp.11-21.

Spinde, T., Rudnitckaia, L., Sinha, K., Hamborg, F., Gipp, B. and Donnay, K. 2021. MBIC--A Media Bias Annotation Dataset Including Annotator Characteristics. *arXiv preprint arXiv:2105.11910*.

Spivak, G.C. 2023. Can the subaltern speak? *Imperialism.* Routledge, pp.171-219.

Stanczak, K. and Augenstein, I. 2021. A survey on gender bias in natural language processing. *arXiv preprint arXiv:2112.14168*.

Standing Working Group on Gender in Research and Innovation. 2018. *Tackling gender bias in research evaluation: Recommendations for action for EU Member States.* [Online]. [Accessed 25 November 2025]. Available from: https://era.gv.at/public/documents/3764/Policy_Brief_en18.pdf

StatCounter. 2023. *Desktop Browser Market Share Worldwide.* [Online].

Steck, H., Ekanadham, C. and Kallus, N. 2024. Is cosine-similarity of embeddings really about similarity? In: *Companion Proceedings of the ACM Web Conference 2024*, pp.887-890.

Steele, C.M. and Aronson, J. 1995. Stereotype threat and the intellectual test performance of African Americans. *Journal of personality and social psychology.* **69**(5), p797.

Stony Brook University. no date. *Biased Language.* [Online]. [Accessed 19 September 2025]. Available from: https://www.stonybrook.edu/commcms/studentaffairs/pledge/resources/biased-language.php

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W. and Wang, W.Y. 2019. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*.

Suresh, H. and Guttag, J. 2021. A framework for understanding sources of harm throughout the machine learning life cycle. In: *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pp.1-9.

Taftahi, J. and Nazari, A. 2017. Quranic Sciences from Abdul Hamid Farahi's Perspective (1863-1930). *International Journal of Applied Linguistics and English Literature.* **6**(7), pp.73-80.

Takahashi, K., Yamamoto, K., Kuchiba, A. and Koyama, T. 2022. Confidence interval for micro-averaged F 1 and macro-averaged F 1 scores. *Applied Intelligence.* **52**(5), pp.4961-4972.

Tan, Z., Li, D., Wang, S., Beigi, A., Jiang, B., Bhattacharjee, A., Karami, M., Li, J., Cheng, L. and Liu, H. 2024. Large language models for data annotation and synthesis: A survey. *arXiv preprint arXiv:2402.13446*.

Tight, M. 2023. Positivity bias in higher education research. *Higher Education Quarterly.* **77**(2), pp.201-214.

Törnberg, P. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*.

Vanmassenhove, E., Hardmeier, C. and Way, A. 2019. Getting gender right in neural machine translation. *arXiv preprint arXiv:1909.05088*.

Voelkel, S., Bates, A., Gleave, T., Larsen, C., Stollar, E.J., Wattret, G. and Mello, L.V. 2023. Lecture capture affects student learning behaviour. *FEBS open bio.* **13**(2), pp.217-232.

Von Denffer, A. 1983. 'Ulūm Al-Qur'ān: An Introduction to the Sciences of the Qur'ān.

Wagner, C., Garcia, D., Jadidi, M. and Strohmaier, M. 2015. It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia. In: *Proceedings of the international AAAI conference on web and social media*, pp.454-463.

Walton, G.M. and Cohen, G.L. 2011. A brief social-belonging intervention improves academic and health outcomes of minority students. *Science.* **331**(6023), pp.1447-1451.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N. and Zhou, M. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems.* **33**, pp.5776-5788.

Wang, X., Kim, H., Rahman, S., Mitra, K. and Miao, Z. 2024. Human-llm collaborative annotation through effective verification of llm labels. In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp.1-21.

Wang, Z.P., Bhandary, P., Wang, Y. and Moore, J.H. 2024. Using GPT-4 to write a scientific review article: a pilot evaluation study. *BioData Mining.* **17**(1), p16.

Wansbrough, J.E. and Rippin, A. 1977. *Quranic studies: sources and methods of scriptural interpretation.* Oxford University Press Oxford.

Warr, M., Pivovarova, M., Mishra, P. and Oster, N.J. 2024. Is ChatGPT racially biased? The case of evaluating student writing. *The Case of Evaluating Student Writing (May 25, 2024)*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V. and Zhou, D. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*. **35**, pp.24824-24837.

Weissburg, I., Anand, S., Levy, S. and Jeong, H. 2024. Llms are biased teachers: Evaluating llm bias in personalized education. *arXiv preprint arXiv:2410.14012*.

Williams, A., Nangia, N. and Bowman, S.R. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Williamson, B. and Eynon, R. 2020. *Historical threads, missing links, and future directions in AI in education*. Taylor & Francis. 45. pp.223-235.

Wineburg, S. 2010. Historical thinking and other unnatural acts. *Phi delta kappan*. **92**(4), pp.81-94.

Yu, J., Zhang, Z., Zhang-li, D., Tu, S., Hao, Z., Li, R.M., Li, H., Wang, Y., Li, H. and Gong, L. 2024. From mooc to maic: Reshaping online teaching and learning through llm-driven agents. *arXiv preprint arXiv:2409.03512*.

Zarrett, N. and Lerner, R.M. 2008. Ways to promote the positive development of children and youth. *Child Trends*. **11**(1), pp.1-5.

Zerfass, A., Verčič, D. and Volk, S.C. 2017. Communication evaluation and measurement: Skills, practices and utilization in European organizations. *Corporate communications: An international journal*. **22**(1), pp.2-18.

Zhang, R., Li, Y., Ma, Y., Zhou, M. and Zou, L. 2023. Llmaaa: Making large language models as active annotators. *arXiv preprint arXiv:2310.19596*.

Zhao, J., Mukherjee, S., Hosseini, S., Chang, K.-W. and Awadallah, A.H. 2020. Gender bias in multilingual embeddings and cross-lingual transfer. *arXiv preprint arXiv:2005.00699*.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V. and Chang, K.-W. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.

Zhao, P. 2003. Mother and I-An analysis of the gender roles of two generations of females in junior middle school English-language teaching materials. *Chinese Education and Society*. **36**(3), pp.27-33.

Zhao, S., Yang, Y., Wang, Z., He, Z., Qiu, L.K. and Qiu, L. 2024. Retrieval augmented generation (rag) and beyond: A comprehensive survey on how to make your llms use external data more wisely. *arXiv preprint arXiv:2409.14924*.

Zhou, K., Ethayarajh, K., Card, D. and Jurafsky, D. 2022. Problems with cosine as a measure of embedding similarity for high frequency words. *arXiv preprint arXiv:2205.05092*.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A. and Fidler, S. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In: *Proceedings of the IEEE international conference on computer vision*, pp.19-27.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z. and Yang, D. 2024. Can large language models transform computational social science? *Computational Linguistics.* **50**(1), pp.237-291.

Zittleman, K. and Sadker, D. 2002. Gender bias in teacher education texts: New (and old) lessons. *Journal of Teacher Education.* **53**(2), pp.168-180.