Exploring Active Learning Algorithms for Data Efficient Language Models

Katerina Margatina

Doctor of Philosophy

Department of Computer Science
University of Sheffield

July 2024

Abstract

Supervised learning is based in the premise that models can effectively solve tasks by learning from numerous examples, mapping inputs to outputs through iterative learning. However, contemporary deep learning models often require vast amounts of labeled data, termed training examples, for optimal performance. Unfortunately, not all training examples contribute equally to the learning process, leading to inefficiencies and resource wastage. Active Learning (AL) has emerged as a powerful paradigm for training language models in a data-efficient manner. By iteratively selecting informative unlabeled data points, which are then annotated by humans to form the training set, AL intelligently guides the training process, optimizing data selection for model improvement over random sampling.

This thesis investigates various aspects of active learning algorithms for language models, focusing on model training, data selection, in-context learning and simulation. The thesis is structured along four key publications that tackle these topics respectively. The first publication addresses the effective adaptation of pretrained language models for AL, highlighting the importance of task-specific fine-tuning. The second publication introduces a novel acquisition function, Contrastive Active Learning (CAL), which selects contrastive examples to improve AL performance. The third publication explores active learning principles for in-context learning with large language models, emphasizing the selection of informative demonstrations for few-shot learning. Lastly, the fourth publication critically examines the limitations of simulating AL experiments and proposes guidelines for future research. Through these contributions, this thesis aims to advance our understanding of AL algorithms for data-efficient language model training.

Lay Summary

Efficiently training language models to understand and generate human-like language is a significant endeavor in artificial intelligence research. Traditionally, this process demands a substantial amount of meticulously labeled data for effective learning. However, such data acquisition can be resource-intensive and time-consuming. Active learning (AL) presents a promising approach to mitigate these challenges by selectively choosing the most informative examples for training, mimicking human learning strategies. This thesis delves into refining AL algorithms tailored for language models, with the aim of enhancing learning efficiency and reducing data requirements.

The research in this thesis is organized around four key studies. The first study introduces novel methodologies to dynamically adapt language models during the learning process, ensuring their adaptability to evolving data distributions. The second study introduces a novel approach to intelligently select hard training examples to optimize the learning process by prioritizing the most informative instances. The third study examines how AL principles can be applied to in-context learning with large language models, focusing on selecting the most useful demonstrations for few-shot learning. The final study evaluates the challenges of simulating AL experiments and offers recommendations for future research.

Overall, this thesis aims to deepen our understanding of active learning algorithms and their potential to train language models more efficiently.

Acknowledgements

My PhD journey over the past four years has been filled with challenges, growth, and joy. Despite moments of difficulty, it has been one of the most rewarding periods of my life. Above all, I'm immensely grateful to my advisor, Nikos, whose support, exceptional mentorship, and expertise have been invaluable. He has always encouraged balance, whether I needed a break, time with family, or sought research opportunities and internships. Nikos is an extraordinary advisor, deeply committed to his students' well-being and development. I feel incredibly fortunate to have had the privilege of working with him.

My time at Sheffield, despite the challenges of the pandemic and lockdowns, has been one of the happiest periods of my life. I'm thankful to my colleagues George Chr., Yida, Mali, Cass, Danae, James, Zeerak, Loic, and Fred for fostering a fantastic culture in our lab. My incredible friends Mara, George Ts., Evianna, Lef, Panos, George P., Tom, Kostis, Mike, Faidonas, and Maria made Sheffield unforgettable. I'll always cherish our hikes, repeated viewings of Twin Peaks, bouldering mishaps, fancy G&Ts with jazz music, and explorations of Sheffield's pub culture.

I spent a significant part of my PhD at the University of Copenhagen, an experience for which I am deeply grateful to Anders for inviting me. My time there was by far the most carefree and invigorating part of this journey. I feel fortunate to have built friendships with remarkable people like Yova, Constanza, Isra, Ilias, Stephanie, Laura, Emanuele, Ruixang, Philip, Heather, Daniel, Desmond, and Stella. I also deeply miss my Copenhagen family—Tolis, Caserita, Vaggelis, and Giorgos B.—and the unforgettable memories we created together in this beautiful city.

I owe a special thanks to my dear friend and academic partner in crime, Giorgos V., whose help has been a guiding force throughout these years. Despite being a coauthor on only one of my papers, his influence permeates all my work. Alexandra, Neli, Christos B., and Efimia are my dream team of brilliant friends, my idols, who have shared their PhD ups and downs with me, and together we have navigated this journey (surprisingly successfuly!).

I want to thank my lifelong friends, Iasonas, Christos K., Foivos, Natalia, and Christiana. Growing up together has been a constant source of strength, and despite following different paths, the love and support we share is rare. Last but not least, I want to thank Thanos, who has been my rock for the past half year, supporting me like no one else during my move to the US and making me a truly happier person.

To close this list of gratitude, I want to extend my deepest appreciation to the most

significant figures in my life—my sister Kelly, my mom, and my dad. Their unwavering support, boundless love, and constant care from afar have been my guiding light, not just through my PhD but in every aspect of my life. Expressing my gratitude fully is beyond words, so I dedicate my PhD to them as a small token of my profound gratitude. Σας ευχαριστώ για όλα.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Katerina Margatina*)

To my extraordinary family, whose unyielding strength and resilience have been a guiding light in navigating the difficulties we have experienced over the recent years.

Contents

Chapter 1

Introduction

Supervised learning operates under the premise that models can effectively tackle tasks by assimilating knowledge from numerous examples, refining their ability to map inputs to outputs through iterative learning. Nonetheless, contemporary deep learning models often demand copious amounts of labeled data, referred to as training examples, to achieve optimal performance. Regrettably, not all training examples contribute equally to the learning process. Active learning (AL) presents a remedy to this challenge by adopting a human-and-model-in-the-loop framework (Cohn et al., 1996; Settles, 2009). In a high-level, the setting includes a human, a pool or stream of unlabeled data and a model (Figure 1.2). An AL algorithm aims to iteratively select the most informative data points from the pool that will then be passed to the human for annotation. Then the acquired labelled data will form the training set that will be used to train the model. The process is typically repeated multiple times until a stopping criterion is met. The goal of this iterative process data efficiency; acquiring the least amount of unlabeled data for training a model that would perform close to the oracle (i.e. the model trained on the entire pool of data). A successful AL algorithm should also enhance model performance beyond random sampling from the pool (Figure 1.1).

The Natural Language Processing (NLP) community has researched active learning algorithms on a plethora of tasks and domains, such as text classification (Ein-Dor et al., 2020; Schröder and Niekler, 2020; Margatina et al., 2022; Schröder et al., 2023), machine translation (Haffari et al., 2009; Dara et al., 2014; Miura et al., 2016; Zhao et al., 2020), named entity recognition (Erdmann et al., 2019; Shen et al., 2017; Wei et al., 2019), natural language inference Snijders et al. (2023), part-of-speech tagging (Chaudhary et al., 2021), coreference (Yuan et al., 2022) and entity resolution (Qian et al., 2017; Kasai et al., 2019), among several others. In our studies we mostly focus on text classification tasks.

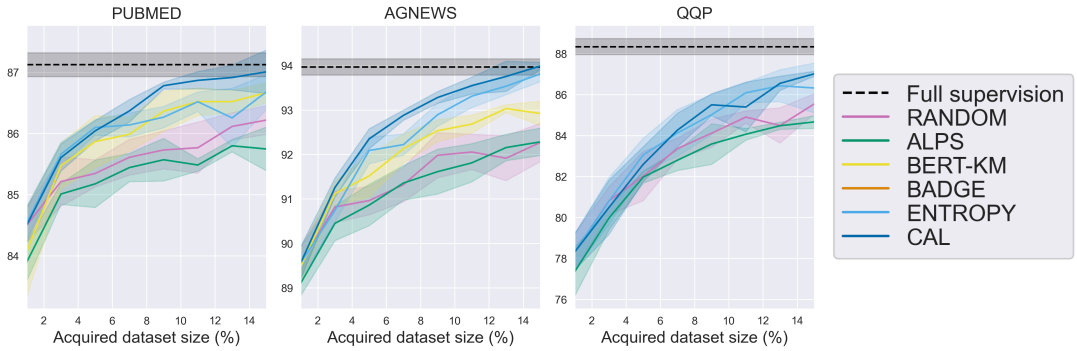


Figure 1.1: Active learning on three datasets (PubMed, AGNews, and qqp), with different acquisition functions (random sampling, Alps, BERT K-means, Badge, Entropy, Cal). Empirical results show that with a small fraction of the training set (from 2% to 15%) a model can reach even the same performance as the model with full supervision (100% of data). Figure from Margatina et al. (2021).

More specifically, in this thesis we focus on the setting of pool-based active learning for text classification tasks. In a pool-based AL setting, an AL algorithm strategically selects the most informative data points from the pool for human annotation, and then a model is trained using the collected labelled data that form the training set. Usually this process is repeated in several rounds. The process is illustrated in Figure 1.2.

Model Training In the model training part of the pipeline we simply train the model with the acquired labeled dataset (Figure 1.2: 2). Interestingly, there are not many studies that explore how we should properly train the model in the low data resource setting of AL. Existing approaches include semi-supervised learning (McCallum and Nigam, 1998; Tomanek and Hahn, 2009; Dasgupta and Ng, 2009; Yu et al., 2022), weak supervision (Ni et al., 2019; Qian et al., 2020; Brantley et al., 2020; Zhang et al., 2022b) and data augmentation (Zhang et al., 2020; Zhao et al., 2020; Hu and Neubig, 2021), with the most prevalent approach currently to be transfer learning from pretrained language models (Ein-Dor et al., 2020; Margatina et al., 2021; Tamkin et al., 2022). We showed large performance gains by adapting the pretrained language model to the task using the unlabeled data of the pool (i.e., task adaptive pretraining by Gururangan et al. (2020)), along with an adaptive fine-tuning technique to account for the varying size of $\mathcal{D}_{\text{lab}}$ (Margatina et al., 2022). In active learning, the initial training dataset often consists of a few tens or hundreds, and is increased until thousands, thus adapting the training strategy is not trivial.

Data Selection The data selection step (Figure 1.2: 4) is probably the core of the AL process and can be performed in various ways. Zhang et al. (2022e) categorize them

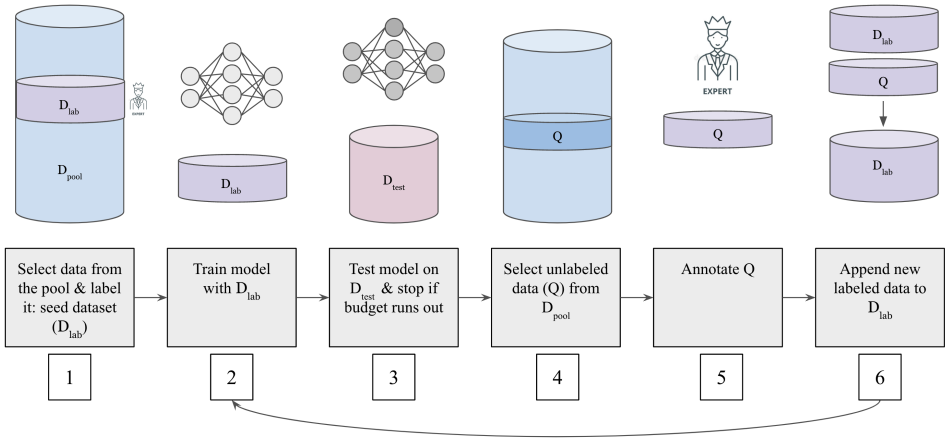


Figure 1.2: High-level overview of the *train-acquire-annotate* steps of the active learning loop. Figure from Margatina and Aletras (2023).

into two main families: informativeness and representativeness. Informativeness-based methods evaluate each candidate instance individually, assigning scores to select the top or bottom instances, with sub-categories including uncertainty sampling (Lewis and Gale, 1994; Culotta and Mccallum, 2005; Zhang and Plank, 2021; Schröder et al., 2022), divergence-based algorithms (Ducoffe and Precioso, 2018; Margatina et al., 2021; Zhang et al., 2022c), disagreement-based Seung et al. (1992); Houlsby et al. (2011); Gal et al. (2017); Siddhant and Lipton (2018); Kirsch et al. (2019); Zeng and Zubiaga (2023), gradient-based (Settles et al., 2007; Settles and Craven, 2008), and performance prediction (Roy and Mccallum, 2001; Konyushkova et al., 2017; Bachman et al., 2017; Liu et al., 2018). Representativeness-based methods, on the other hand, consider the correlation between instances to avoid sampling bias, with sub-categories such as density-based methods (Ambati et al., 2010; Zhao et al., 2020; Zhu et al., 2008) and batch diversity approaches (Gissin and Shalev-Shwartz, 2019; Erdmann et al., 2019), where core-set is the most common (Sener and Savarese, 2018). Additionally, there are hybrid methods that combine both informativeness and representativeness (Brinker, 2003; Bodó et al., 2011; Zhu et al., 2008; Geifman and El-Yaniv, 2017; Zhdanov, 2019; Yu et al., 2022). Despite the variety of methods, no single acquisition function consistently outperforms others, making the choice of data acquisition a continuing research focus. Selecting the most useful data for annotation through active learning can be critical to achieve high test set performance while opting for data efficiency. Figure 1.3 illustrates the considerably different performance curves with various AL data acquisition strategies, with some achieving the full dataset performance (i.e., if using 100% of the training data) with less than 15% of the dataset, while others can have detrimental effects and perform similarly with both 1% and 15% of the data.

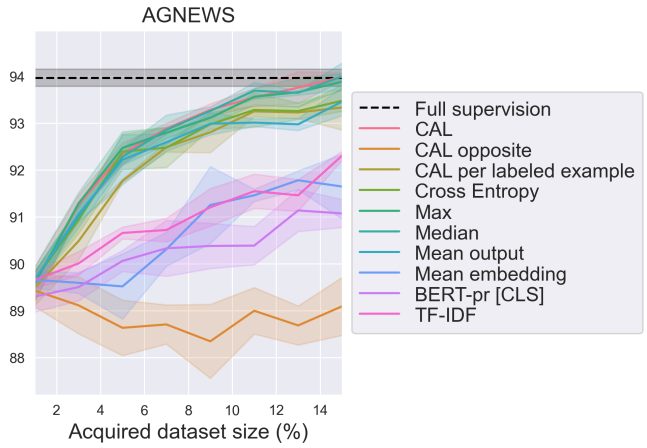


Figure 1.3: Different data selection methods can lead to considerably different performance in the held-out test set when different labeled datasets are acquired through active learning. Figure from Margatina et al. (2021).

Data Annotation Finally, the acquired unlabeled data is sent to humans for annotation (Figure 1.2: 5). In simulation settings, researchers often overlook this step because the labels are already known (i.e., in simulation labeled datasets are treated us unlabeled). However, not all examples are equally easy to annotate; difficult instances for classifiers are typically hard for humans too (Hachey et al., 2005; Baldridge and Osborne, 2004). This discrepancy suggests that current experimental settings are limited and highlights the need for cost-aware selection strategies, which consider the varying difficulty of instances and the expertise of annotators (Donmez and Carbonell, 2008; Baldridge and Palmer, 2009; Tomanek and Hahn, 2010; Wei et al., 2019).

1.1 Research Aims and Objectives

Overall, we lay out four desiderata that will be addressed by the approaches proposed in this thesis:

Developing model training algorithms that are effective throughout active learning iterations This research study is motivated by the observation that off-the-shelf pretrained language models are not effectively adapted to downstream tasks during active learning in natural language processing (NLP). Addressing this issue is important as it aims to improve the data efficiency of AL by proposing a method to better adapt and fine-tune LMs, ensuring robust performance in both low and high resource scenarios, ultimately enhancing the efficacy of AL strategies.

Strategically selecting the most useful unlabeled data The motivation for this part of the thesis is the need to improve active learning strategies by combining the

strengths of uncertainty and diversity sampling. The study introduces a novel acquisition function, Contrastive Active Learning (CAL), which selects data points that are similar in the model feature space but have maximally different predictive likelihoods. The importance of this research lies in its demonstrated ability to outperform existing methods across multiple natural language understanding tasks and datasets, offering a more effective balance between uncertainty and diversity in AL.

Exploring active learning principles beyond supervised learning The drive behind this work is the need to optimize the selection of demonstrations for few-shot learning in large language models (LLMs), an area that has received limited attention previously. This study addresses this gap by approaching demonstration selection as a pool-based Active Learning problem of a single iteration, and finds that similarity-based selection consistently outperforms other methods. The importance of this work lies in its extensive experimentation and analysis, showing that using semantically similar demonstrations significantly enhances performance, even outperforming much larger models using random demonstrations, thus underscoring the critical role of demonstration selection in few-shot learning.

Addressing pain points in active learning simulation Only a limited body of research has delved into the pain points of AL. For example, studies have revealed that AL algorithms tend to acquire collective outliers, leading to the failure of several AL approaches to surpass random sampling in certain tasks. Moreover, the benefits of AL may not generalize reliably across different models and domains, highlighting the need for a nuanced understanding of its limitations.

1.2 Thesis Overview: Publications and Contributions

This section lists the contributions made throughout this thesis. It follows a thesis by publications format and consists of a collection of four papers where each paper corresponds to an individual chapter.

Chapter 2 introduces two techniques for fine-tuning pre-trained language models in a low-data resource setting of active learning and it is based on the publication *"Effectively Adapting Pretrained Language Models for Active Learning"* (Margatina et al., 2022).

Chapter 3 introduces Contrastive Active Learning (CAL), a novel acquisition function for active learning, and it is based on the publication *"Active Learning by Acquiring Contrastive Examples"* (Margatina et al., 2021).

Chapter 4 explores how active learning algorithms can be applied to select demon-

strations for in-context learning and it is based on the publication *"Active Learning Principles for In-Context Learning"* (Margatina et al., 2023).

Chapter 5 criticizes common practises in active learning simulation experiments and proposes guidelines for future work on the field. It is based on the publication *"On the Limitations of Simulating Active Learning"* (Margatina and Aletras, 2023).

Chapter 6 finally contains our conclusion where we summarize our findings and provide an outlook into the future.

1.3 Published Work

The work in this dissertation primarily relates to the following peer-reviewed articles:

1. **Margatina, K.**, Vernikos, G., Barrault, L., Aletras, N. (2021). Active Learning by Acquiring Contrastive Examples. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2021

2. **Margatina, K.**, Barrault, L., Aletras, N. (2022). Effectively Adapting Pretrained Language Models for Active Learning. *In Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2022

3. **Margatina, K.**, Schick, T., Aletras, N., Dwivedi-Yu, J. (2023). Active Learning Principles for In-Context Learning with Large Language Models. *In Findings of the Association for Computational Linguistics*, 2023

4. **Margatina, K.** and Aletras, N. (2023). On the Limitations of Simulating Active Learning. *In Findings of the Association for Computational Linguistics*, 2023

The following article is related, but will not be extensively discussed in this thesis:

5. Snijders, A., Kiela D., **Margatina, K.** (2023). Investigating Multi-source Active Learning for Natural Language Inference. *In Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, 2023

Finally, while not directly related, the following articles have also been completed over the course of the PhD:

6. Vernikos G., **Margatina, K.**, Chronopoulou, A., Androutsopoulos, I. (2020). Domain Adversarial Fine-Tuning as an Effective Regularizer. *In Findings of the Association for Computational Linguistics*, 2020

7. Yamaguchi, A., Chrysostomou, G., **Margatina, K.**, Aletras, N. (2021). Frustratingly simple pretraining alternatives to masked language modeling. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2021

8. Hershcovich, D., Frank, S., Lent, H., Lhoneux, M., Abdou, M., Brandl, S., Bugliarello, E., Cabello Piqueras, L., Chalkidis, I., Cui, R., Fierro, C., **Margatina, K.**, Rust, P., Søgaard, A. (2022). Challenges and strategies in cross-cultural NLP. *In Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2022

9. **Margatina, K.**, Wang, S., Vyas, Y., Anna John, N., Benajiba, Y., Ballesteros, M. (2023). Dynamic Benchmarking of Masked Language Models on Temporal Concept Drift with Multiple Views. *In Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, 2023

10. Alajrami, A., **Margatina, K.**, Aletras, N. (2023). Understanding the Role of Input Token Characters in Language Models: How Does Information Loss Affect Performance?. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2023

11. Kirk, H., Whitefield, A., Röttger, P., Bean, A., **Margatina, K.**, Ciro, J., Mosquera, R., Bartolo, M., Williams, A., He, H., Vidgen, B., Hale, S. (2024). The PRISM Alignment Project: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models. *Arxiv*, 2024

Chapter 2

Publication I: Effectively Adapting Language Models for Active Learning

The main contribution of this chapter is the paper *On The Importance of Effectively Adapting Pretrained Language Models for Active Learning*, which was published at the *Annual Meeting of the Association for Computational Linguistics* in May 2022. We first outline the motivation (Section 2.1), followed by the paper itself (Section 2.2), the impact that it has had so far (Section 2.3), and discussion (Section 2.4).

2.1 Introduction

The work presented in this section was largely inspired by the task adaptive pretraining paradigm (Gururangan et al., 2020), which at the time of its introduction, was the most successful approach in leveraging unlabelled data to better adapt a pretrained model, such as BERT (Devlin et al., 2019), to the task-specific domain. The method is quite simple; collect task-specific corpora and continue pretraining the model with the language modelling loss for a few training steps and after convergence fine-tune it to the task. We thought that this would be the ideal setting for a pool-based active learning setting, where we have already access to a large pool of unlabeled data. Before we start AL we first do task-adaptive pretraining (TAPT) of the model for 100K training steps. We use the new model checkpoint to initialize our first model for pool-based AL which we then fined-tuned with the acquired labeled training dataset using fixed hyperparameters proposed by Devlin et al. (2019), that include a fixed number of 3 training epochs, learning rate warmup over the first 10% of the steps and AdamW

optimizer without bias correction. After early experimentation we found that TAPT was highly successful, but we thought that there was still room for improvement. At that point, we were fine-tuning the model at each AL iteration the training dataset that gradually increased in size (as more labeled data was acquired through AL). We realized that we had to further adapt our approach in order to take into account the change in the size of the actively acquired dataset that started from a few hundreds of examples (low-data resource) to thousands (high-data resource). We experimented we several adaptations and finalized our training approach (FT+) at using early stopping with 20 epochs based on the validation loss, `min(`10% `of total steps,` 100`)`, learning rate $2e-5$, bias correction and 5 evaluation steps per epoch. Section 2 of the paper presents our proposed methodology and Appendix A3-A4 the training details.

We found that our two contributions (TAPT and FT+) to the standard pool-based AL training pipeline had a notable effect in improving AL performance, as demonstrated by our experiments in Section 3 of the paper. For instance, we found that using AL with entropy acquisition function and our methodology we managed to get close to the full model performance (i.e., the score in the test set if the model was trained with all available training data) using less than 10% of the training data. Our analysis, shown in Section 4 of the paper, further validates that each component of our proposed methodology was crucial in order to provide an approach that leverages the full potential of the BERT AL model.

2.2 The Paper

Author Contributions

The paper is co-authored by myself, Loïc Barrault and Nikolaos Aletras. Nikolaos Aletras seeded the idea to explore ways to use BERT to boost AL performance, supervised the project, offered suggestions, and helped write the paper. As the lead author, I developed the proposed methodology, performed the experiments, and wrote the paper. Loïc Barrault helped with the identifying limitations at various stages of the development of the methodology, participated in various discussions and proofread the paper.

On the Importance of Effectively Adapting Pretrained Language Models for Active Learning

Katerina Margatina[♠] **Loïc Barrault**[♣] **Nikolaos Aletras**[♠]
[♠]University of Sheffield, [♣]University of Le Mans
{k.margatina,n.aletras}@sheffield.ac.uk
loic.barrault@univ-lemans.fr

Abstract

Recent Active Learning (AL) approaches in Natural Language Processing (NLP) proposed using off-the-shelf pretrained language models (LMs). In this paper, we argue that these LMs are not adapted effectively to the downstream task during AL and we explore ways to address this issue. We suggest to first adapt the pretrained LM to the target task by continuing training with all the available *unlabeled* data and then use it for AL. We also propose a simple yet effective fine-tuning method to ensure that the adapted LM is properly trained in both low and high resource scenarios during AL. Our experiments demonstrate that our approach provides substantial data efficiency improvements compared to the standard fine-tuning approach, suggesting that a poor training strategy can be catastrophic for AL.[1]

1 Introduction

Active Learning (AL) is a method for training supervised models in a data-efficient way (Cohn et al., 1996; Settles, 2009). It is especially useful in scenarios where a large pool of unlabeled data is available but only a limited annotation budget can be afforded; or where expert annotation is prohibitively expensive and time consuming. AL methods iteratively alternate between (i) model training with the labeled data available; and (ii) data selection for annotation using a stopping criterion, e.g. until exhausting a fixed annotation budget or reaching a pre-defined performance on a held-out dataset.

Data selection is performed by an acquisition function that ranks unlabeled data points by some *informativeness* metric aiming to improve over random selection, using either uncertainty (Lewis and Gale, 1994; Cohn et al., 1996; Gal et al., 2017; Kirsch et al., 2019; Zhang and Plank, 2021), diversity (Brinker, 2003; Bodó et al., 2011; Sener and Savarese, 2018), or both (Ducoffe and Precioso, 2018; Ash et al., 2020; Yuan et al., 2020; Margatina et al., 2021).

Previous AL approaches in NLP use task-specific neural models that are trained from scratch at each iteration (Shen et al., 2017; Siddhant and Lipton, 2018; Prabhu et al., 2019; Ikhwantri et al., 2018; Kasai et al., 2019). However, these models are usually outperformed by pretrained language models (LMs) adapted to end-tasks (Howard and Ruder, 2018), making them suboptimal for AL. Only recently, pretrained LMs such as BERT (Devlin et al., 2019) have been introduced in AL settings (Yuan et al., 2020; Ein-Dor et al., 2020; Shelmanov et al., 2021; Karamcheti et al., 2021; Margatina et al., 2021). Still, they are trained at each AL iteration with a standard fine-tuning approach that mainly includes a pre-defined number of training epochs, which has been demonstrated to be unstable, especially in small datasets (Zhang et al., 2020; Dodge et al., 2020; Mosbach et al., 2021). Since AL includes both low and high data resource settings, the AL model training scheme should be robust in both scenarios.[2]

To address these limitations, we introduce a suite of effective training strategies for AL (§2). Contrary to previous work (Yuan et al., 2020; Ein-Dor et al., 2020; Margatina et al., 2021) that also use BERT (Devlin et al., 2019), our proposed method accounts for various data availability settings and the instability of fine-tuning. First, we continue *pretraining* the LM with the available *unlabeled* data to adapt it to the task-specific domain. This way, we leverage not only the available labeled data at each AL iteration, but the entire unlabeled pool. Second, we further propose a simple yet effective fine-tuning method that is robust in both low and high resource data settings for AL.

[1]For all experiments in this paper, we have used the code provided by Margatina et al. (2021): https://github.com/mourga/contrastive-active-learning

[2]During the first few AL iterations the available labeled data is limited (*low-resource*), while it could become very large towards the last iterations (*high-resource*).

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 825 - 836

We explore the effectiveness of our approach on five standard natural language understandings tasks with various acquisition functions, showing that it outperforms all baselines (§3). We also conduct an analysis to demonstrate the importance of effective adaptation of pretrained models for AL (§4). Our findings highlight that the LM adaptation strategy can be more critical than the actual data acquisition strategy.

2 Adapting & Fine-tuning Pretrained Models for Active Learning

Given a downstream classification task with C classes, a typical AL setup consists of a pool of unlabeled data $\mathcal{D}_{\textbf{pool}}$, a model $\mathcal{M}$, an annotation budget b of data points and an acquisition function $a(.)$ for selecting k unlabeled data points for annotation (i.e. acquisition size) until b runs out. The AL performance is assessed by training a model on the actively acquired dataset and evaluating on a held-out test set $\mathcal{D}_{\textbf{test}}$.

Adaptation (TAPT) Inspired by recent work on transfer learning that shows improvements in downstream classification performance by continuing the pretraining of the LM with the task data (Howard and Ruder, 2018) we add an extra step to the AL process by continuing pretraining the LM (i.e. Task-Adaptive Pretraining TAPT), as in Gururangan et al. (2020). Formally, we use an LM, such as BERT (Devlin et al., 2019), $\mathcal{P}(x; W_0)$ with weights W_0, that has been already pretrained on a large corpus. We fine-tune $\mathcal{P}(x; W_0)$ with the available unlabeled data of the downstream task $\mathcal{D}_{\textbf{pool}}$, resulting in the task-adapted LM $\mathcal{P}_{\text{TAPT}}(x; W_0')$ with new weights W_0' (cf. line 2 of algorithm 1).

Fine-tuning (FT+) We now use the adapted LM $\mathcal{P}_{\text{TAPT}}(x; W_0')$ for AL. At each iteration i, we initialize our model $\mathcal{M}_i$ with the pretrained weights W_0' and we add a task-specific feedforward layer for classification with weights W_c on top of the `[CLS]` token representation of BERT-based $\mathcal{P}_{\text{TAPT}}$. We fine-tune the classification model $\mathcal{M}_i(x; [W_0', W_c])$ with all $x \in \mathcal{D}_{\textbf{lab}}$. (cf. line 6 to 8 of algorithm 1).

Recent work in AL (Ein-Dor et al., 2020; Yuan et al., 2020) uses the standard fine-tuning method proposed in Devlin et al. (2019) which includes a fixed number of 3 training epochs, learning rate warmup over the first 10% of the steps and AdamW optimizer (Loshchilov and Hutter, 2019) without bias correction, among other hyperparameters.

Algorithm 1: AL with Pretrained LMs

Input: unlabeled data $\mathcal{D}_{\textbf{pool}}$, pretrained LM $\mathcal{P}(x; W_0)$, acquisition size k, AL iterations T, acquisition function a

1 $\mathcal{D}_{\textbf{lab}} \leftarrow \emptyset$
2 $\mathcal{P}_{\text{TAPT}}(x; W_0') \leftarrow$ Train $\mathcal{P}(x; W_0)$ on $\mathcal{D}_{\textbf{pool}}$
3 $\mathcal{Q}_0 \leftarrow \text{RANDOM}(.), |\mathcal{Q}_0| = k$
4 $\mathcal{D}_{\textbf{lab}} = \mathcal{D}_{\textbf{lab}} \cup \mathcal{Q}_0$
5 $\mathcal{D}_{\textbf{pool}} = \mathcal{D}_{\textbf{pool}} \setminus \mathcal{Q}_0$
6 **for** $i \leftarrow 1$ **to** T **do**
7 $\quad \mathcal{M}_i(x; [W_0', W_c]) \leftarrow$ Initialize from $\mathcal{P}_{\text{TAPT}}(x; W_0')$
8 $\quad \mathcal{M}_i(x; W_i) \leftarrow$ Train model on $\mathcal{D}_{\textbf{lab}}$
9 $\quad \mathcal{Q}_i \leftarrow a(\mathcal{M}_i, \mathcal{D}_{\textbf{pool}}, k)$
10 $\quad \mathcal{D}_{\textbf{lab}} = \mathcal{D}_{\textbf{lab}} \cup \mathcal{Q}_i$
11 $\quad \mathcal{D}_{\textbf{pool}} = \mathcal{D}_{\textbf{pool}} \setminus \mathcal{Q}_i$
12 **end**

Output: $\mathcal{D}_{\textbf{lab}}$

We follow a different approach by taking into account insights from few-shot fine-tuning literature (Mosbach et al., 2021; Zhang et al., 2020; Dodge et al., 2020) that proposes longer fine-tuning and more evaluation steps during training. [3] We combine these guidelines to our fine-tuning approach by using early stopping with 20 epochs based on the validation loss, learning rate $2e-5$, bias correction and 5 evaluation steps per epoch. However, increasing the number of epochs from 3 to 20, also increases the warmup steps (10% of total steps[4]) almost 7 times. This may be problematic in scenarios where the dataset is large but the optimal number of epochs may be small (e.g. 2 or 3). To account for this limitation in our AL setting where the size of training set changes at each iteration, we propose to select the warmup steps as $min(10\% \text{ of total steps}, 100)$. We denote standard fine-tuning as SFT and our approach as FT+.

3 Experiments & Results

Data We experiment with five diverse natural language understanding tasks: question classification

[3] In this paper we use *few-shot* to describe the setting where there are *few* labeled data available and therefore *few-shot fine-tuning* corresponds to fine-tuning a model on limited labeled training data. This is different than the few-shot setting presented in recent literature (Brown et al., 2020), where no model weights are updated.

[4] Some guidelines propose an even smaller number of warmup steps, such as 6% in RoBERTa (Liu et al., 2020).

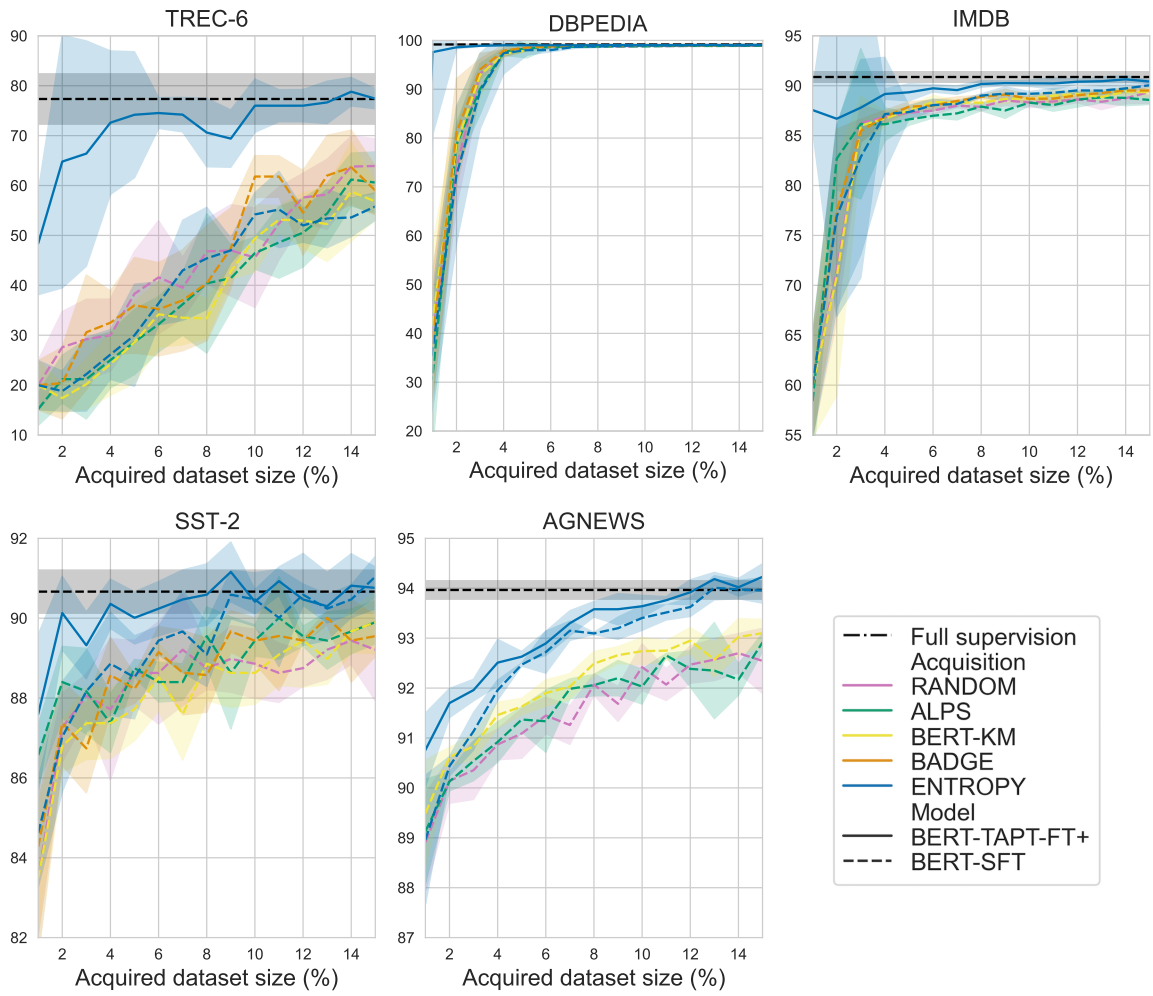


Figure 1: Test accuracy during AL iterations. We plot the median and standard deviation across five runs.

DATASETS	TRAIN	VAL	TEST	k	C
TREC-6	4.9K	546	500	1%	6
DBPEDIA	20K	2K	70K	1%	14
IMDB	22.5K	2.5K	25K	1%	2
SST-2	60.6K	6.7K	871	1%	2
AGNEWS	114K	6K	7.6K	0.5%	4

Table 1: Datasets statistics for $\mathcal{D}_{\text{pool}}$, $\mathcal{D}_{\text{val}}$ and $\mathcal{D}_{\text{test}}$ respectively. k stands for the acquisition size (% of $\mathcal{D}_{\text{pool}}$) and C the number of classes.

(TREC-6; Voorhees and Tice (2000)), sentiment analysis (IMDB; Maas et al. (2011), SST-2 Socher et al. (2013)) and topic classification (DBPEDIA, AGNEWS; Zhang et al. (2015)), including binary and multi-class labels and varying dataset sizes (Table 1). More details can be found in Appendix A.1.

Experimental Setup We perform all AL experiments using `BERT-base` (Devlin et al., 2019) and ENTROPY, BERTKM, ALPS (Yuan et al., 2020), BADGE (Ash et al., 2020) and RANDOM (baseline) as the acquisition functions. We pair our proposed training approach TAPT-FT+ with ENTROPY acquisition. We refer the reader to Appendix A for an extended description of our experimental setup, including the datasets used (§A.1), the training and AL details (§A.2), the model hyperparameters (§A.3) and the baselines (§A.4).

Results Figure 1 shows the test accuracy during AL iterations. We first observe that our proposed approach (TAPT-FT+) achieves large data efficiency reaching the full-dataset performance within the 15% budget for all datasets, in contrast to the standard AL approach (BERT-SFT). The effectiveness of our approach is mostly notable in the smaller datasets. In TREC-6, it achieves the goal accuracy with almost 10% annotated data, while in DBPEDIA only in the first iteration with 2% of the data. After the first AL iteration in IMDB, TAPT-FT+, it achieves only 2.5 points of accuracy lower than the

performance when using 100% of the data. In the larger SST-2 and AGNEWS datasets, it is closer to the baselines but still outperforms them, achieving the full-dataset performance with 8% and 12% of the data respectively. We also observe that in all five datasets, the addition of our proposed pretraining step (TAPT) and fine-tuning technique (FT+) leads to large performance gains, especially in the first AL iterations. This is particularly evident in TREC-6, DBPEDIA and IMDB datasets, where after the *first* AL iteration (i.e. equivalent to 2% of training data) TAPT+FT+ with ENTROPY is 45, 30 and 12 points in accuracy higher than the ENTROPY baseline with BERT and SFT.

Training vs. Acquisition Strategy We finally observe that the performance curves of the various acquisition functions considered (i.e. dotted lines) are generally close to each other, suggesting that the choice of the acquisition strategy may not affect substantially the AL performance in certain cases. In other words, we conclude that *the training strategy can be more important than the acquisition strategy*. We find that uncertainty sampling with ENTROPY is generally the best performing acquisition function, followed by BADGE.[5] Still, finding a universally well-performing acquisition function, independent of the training strategy, is an open research question.

4 Analysis & Discussion

4.1 Task-Adaptive Pretraining

We first present details of our implementation of TAPT (§2) and reflect on its effectiveness in the AL pipeline. Following Gururangan et al. (2020), we continue pretraining BERT for the MLM task using all the unlabeled data $\mathcal{D}_{\textbf{pool}}$ for all datasets separately. We plot the learning curves of BERT-TAPT for all datasets in Figure 2. We first observe that the masked LM loss is steadily decreasing for DBPEDIA, IMDB and AGNEWS across optimization steps, which correlates with the high early AL performance gains of TAPT in these datasets (Fig. 1). We also observe that the LM overfits in TREC-6 and SST-2 datasets. We attribute this to the very small training dataset of TREC-6 and the informal textual style of SST-2. Despite the fact that the SST-2 dataset includes approximately 67K of training data, the sentences are very short (i.e. average length of 9.4 words per sentence). We hypothesize the LM overfits because of the lack of long and more diverse sentences. We provide more details on TAPT at the Appendix B.1.

[5] We provide results with additional acquisition functions in the Appendix B.2 and B.3.

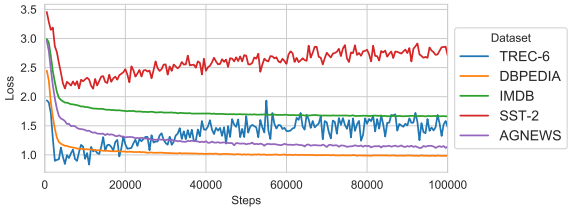


Figure 2: Validation MLM loss during TAPT.

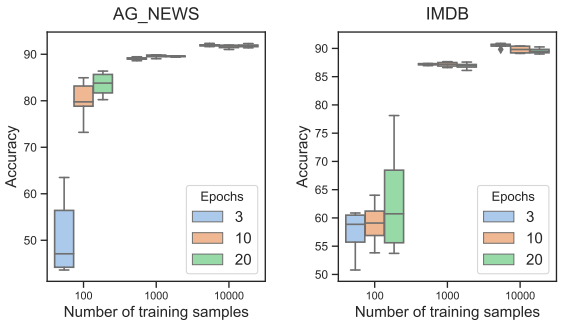


Figure 3: Few-shot standard BERT fine-tuning.

4.2 Few-shot Fine-tuning

In this set of experiments, we aim to highlight that it is crucial to consider the few-shot learning problem in the early AL stages, which is often neglected in literature. This is more important when using pretrained LMs, since they are overparameterized models that require adapting their training scheme in low data settings to ensure robustness.

To illustrate the potential ineffectiveness of standard fine-tuning (SFT), we randomly undersample the AGNEWS and IMDB datasets to form low, medium and high resource data settings (i.e. 100, 1,000 and 10,000 training samples), and train BERT for a fixed number of 3, 10, and 20 epochs. We repeat this process with 10 different random seeds to account for stochasticity in sampling and we plot the test accuracy in Figure 3. Figure 3 shows that SFT is suboptimal for low data settings (e.g. 100 samples), indicating that more optimization steps (i.e. epochs) are needed for the model to adapt to the few training samples (Zhang et al., 2020; Mosbach et al., 2021). As the training samples increase (e.g. 1,000), fewer epochs are often better. It is thus evident that there is not a clearly optimal way to choose a predefined number

of epochs to train the model given the number of training examples. This motivates the need to find a fine-tuning policy for AL that effectively adapts to the data resource setting of each iteration (independently of the number of training examples or dataset), which is mainly tackled by our proposed fine-tuning approach FT+ (§2).

4.3 Ablation Study

We finally conduct an ablation study to evaluate the contribution of our two proposed steps to the AL pipeline; the pretraining step (TAPT) and fine-tuning method (FT+). We show that the addition of both methods provides large gains compared to standard fine-tuning (SFT) in terms of accuracy, data efficiency and uncertainty calibration. We compare BERT with SFT, BERT with FT+ and BERT-TAPT with FT+. Along with test accuracy, we also evaluate each model using uncertainty estimation metrics (Ovadia et al., 2019): Brier score, negative log likelihood (NLL), expected calibration error (ECE) and entropy. A well-calibrated model should have high accuracy and low uncertainty.

Figure 4 shows the results for the smallest and largest datasets, TREC-6 and AGNEWS respectively. For TREC-6, training BERT with our fine-tuning approach FT+ provides large gains both in accuracy and uncertainty calibration, showing the importance of fine-tuning the LM for a larger number of epochs in low resource settings. For the larger dataset, AGNEWS, we see that BERT with SFT performs equally to FT+ which is the ideal scenario. We see that our fine-tuning approach does not deteriorate the performance of BERT given the large increase in warmup steps, showing that our simple strategy provides robust results in both high and low resource settings. After demonstrating that FT+ yields better results than SFT, we next compare BERT-TAPT-FT+ against BERT-FT+. We observe that in both datasets BERT-TAPT outperforms BERT, with this being particularly evident in the early iterations. This confirms our hypothesis that by implicitly using the entire pool of unlabeled data for extra pretraining (TAPT), we boost the performance of the AL model using less data.

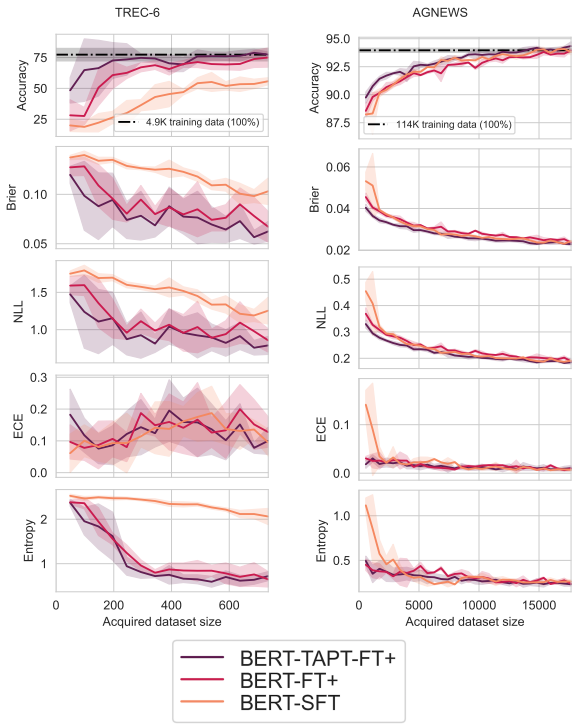


Figure 4: Ablation study for TAPT and FT+.

5 Conclusion

We have presented a simple yet effective training scheme for AL with pretrained LMs that accounts for varying data availability and instability of fine-tuning. Specifically, we propose to first continue pretraining the LM with the available unlabeled data to *adapt* it to the task-specific domain. This way, we leverage not only the available labeled data at each AL iteration, but the entire unlabeled pool. We further propose a method to *fine-tune* the model during AL iterations so that training is robust in both low and high resource data settings.

Our experiments show that our approach yields substantially better results than standard fine-tuning in five standard NLP datasets. Furthermore, we find that *the training strategy can be more important than the acquisition strategy*. In other words, a poor training strategy can be a crucial impediment to the effectiveness of a good acquisition function, and thus limit its effectiveness (even over random sampling). Hence, our work highlights how critical it is to properly adapt a pretrained LM to the low data resource AL setting.

As state-of-the-art models in NLP advance rapidly, in the future we would be interested in exploring the use of larger LMs, such as GPT-3 (Brown et al., 2020) and FLAN (Wei et al., 2022). These models have achieved impressive performance in very low data resource settings (e.g. zero-shot and few-shot), so we would imagine they would be good candidates for the challenging setting of active learning.

Acknowledgments

We would like to thank Giorgos Vernikos, our colleagues at the Sheffield NLP group for feedback on an earlier version of this paper, and all the anonymous reviewers for their constructive comments. KM and NA are supported by Amazon through the Alexa Fellowship scheme.

References

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2020. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*.

Zalán Bodó, Zsolt Minier, and Lehel Csató. 2011. Active learning with clustering. In *Proceedings of the Active Learning and Experimental Design workshop In conjunction with AISTATS 2010*, volume 16, pages 127–139.

Klaus Brinker. 2003. Incorporating diversity in active learning with support vector machines. In *Proceedings of the International Conference on Machine Learning*, pages 59–66.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. 1996. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4(1):129–145.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah A. Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *ArXiv*.

Melanie Ducoffe and Frederic Precioso. 2018. Adversarial active learning for deep networks: a margin based approach.

Liat Ein-Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2020. Active learning for BERT: An empirical study. In *Proceedings of theConference on Empirical Methods in Natural Language Processing*, pages 7949–7962.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the International Conference on Machine Learning*, volume 48, pages 1050–1059.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep Bayesian active learning with image data. In *Proceedings of the International Conference on Machine Learning*, volume 70, pages 1183–1192.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian active learning for classification and preference learning. *ArXiv*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 328–339.

Fariz Ikhwantri, Samuel Louvan, Kemal Kurniawan, Bagas Abisena, Valdi Rachman, Alfan Farizki Wicaksono, and Rahmad Mahendra. 2018. Multi-task active learning for neural semantic role labeling on low resource conversational corpus. In *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*, pages 43–50.

Siddharth Karamcheti, Ranjay Krishna, Li Fei-Fei, and Christopher Manning. 2021. Mind your outliers! investigating the negative impact of outliers on active learning for visual question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7265–7281, Online. Association for Computational Linguistics.

Jungo Kasai, Kun Qian, Sairam Gurajada, Yunyao Li, and Lucian Popa. 2019. Low-resource deep entity resolution with transfer and active learning. In *Proceedings of the Conference of the Association for Computational Linguistic*, pages 5851–5861.

Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. 2019. BatchBALD: Efficient and diverse batch acquisition for deep bayesian active learning. In *Neural Information Processing Systems*, pages 7026–7037.

David D. Lewis and William A. Gale. 1994. A sequential algorithm for training text classifiers. In *In Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Ro{bert}a: A robustly optimized {bert} pretraining approach.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

David Lowell and Zachary C Lipton. 2019. Practical obstacles to deploying active learning. *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, pages 21–30.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150.

Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. 2021. Active learning by acquiring contrastive examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 650–663, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the stability of fine-tuning {bert}: Misconceptions, explanations, and strong baselines. In *International Conference on Learning Representations*.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, pages 13991–14002.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035.

Ameya Prabhu, Charles Dognin, and Maneesh Singh. 2019. Sampling bias in deep active classification: An empirical study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, pages 4056–4066.

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*.

Burr Settles. 2009. Active learning literature survey. Computer sciences technical report.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*.

Artem Shelmanov, Dmitri Puzyrev, Lyubov Kupriyanova, Denis Belyakov, Daniil Larionov, Nikita Khromov, Olga Kozlova, Ekaterina Artemova, Dmitry V. Dylov, and Alexander Panchenko. 2021. Active learning for sequence tagging with deep pre-trained models and Bayesian uncertainty estimates. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1698–1712, Online. Association for Computational Linguistics.

Yanyao Shen, Hyokun Yun, Zachary Lipton, Yakov Kronrod, and Animashree Anandkumar. 2017. Deep active learning for named entity recognition. In *Proceedings of the Workshop on Representation Learning for NLP*, pages 252–256.

Aditya Siddhant and Zachary C Lipton. 2018. Deep bayesian active learning for natural language processing: Results of a Large-Scale empirical study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 2904–2909.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642.

N Srivastava, G Hinton, A Krizhevsky, and others. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Ellen Voorhees and Dawn Tice. 2000. The trec-8 question answering track evaluation. *Proceedings of the Text Retrieval Conference*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Michelle Yuan, Hsuan-Tien Lin, and Jordan Boyd-Graber. 2020. Cold-start active learning through self-supervised language modeling.

Mike Zhang and Barbara Plank. 2021. Cartography active learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 395–406, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q. Weinberger, and Yoav Artzi. 2020. Revisiting few-sample bert fine-tuning. *ArXiv*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28, pages 649–657. Curran Associates, Inc.

A Appendix: Experimental Setup

A.1 Datasets

We experiment with five diverse natural language understanding tasks including binary and multi-class labels and varying dataset sizes (Table 1). The first task is question classification using the six-class version of the small TREC-6 dataset of open-domain, fact-based questions divided into broad semantic categories (Voorhees and Tice, 2000). We also evaluate our approach on sentiment analysis using the binary movie review IMDB dataset (Maas et al., 2011) and the binary version of the SST-2 dataset (Socher et al., 2013). We finally use the large-scale AGNEWS and DBPEDIA datasets from Zhang et al. (2015) for topic classification. We undersample the latter and form a $\mathcal{D}_{\textbf{pool}}$ of 20K examples and $\mathcal{D}_{\textbf{val}}$ 2K as in Margatina et al. (2021). For TREC-6, IMDB and SST-2 we randomly sample 10% from the training set to serve as the validation set, while for AGNEWS we sample 5%. For the DBPEDIA dataset we undersample both training and validation datasets (from the standard splits) to facilitate our AL simulation (i.e. the original dataset consists of 560K training and 28K validation data examples). For all datasets we use the standard test set, apart from the SST-2 dataset that is taken from the GLUE benchmark (Wang et al., 2019) we use the development set as the held-out test set (and subsample a development set from the original training set).

A.2 Training & AL Details

We use BERT-BASE (Devlin et al., 2019) and fine-tune it (TAPT §2) for 100K steps, with learning rate $2e-05$ and the rest of hyperparameters as in Gururangan et al. (2020) using the HuggingFace library (Wolf et al., 2020). We evaluate the model 5 times per epoch on $\mathcal{D}_{val}$ and keep the one with the lowest validation loss as in Dodge et al. (2020). We use the code provided by Kirsch et al. (2019) for the uncertainty-based acquisition functions and Yuan et al. (2020) for ALPS, BADGE and BERTKM. We use the standard splits provided for all datasets, if available, otherwise we randomly sample a validation set. We test all models on a held-out test set. We repeat all experiments with five different random seeds resulting into different initializations of $\mathcal{D}_{\textbf{lab}}$ and the weights of the extra task-specific output feedforward layer. For all datasets we use as budget the 15% of $\mathcal{D}_{\textbf{pool}}$. Each experiment is run on a single Nvidia Tesla V100 GPU.

A.3 Hyperparameters

For all datasets we train BERT-BASE (Devlin et al., 2019) from the HuggingFace library (Wolf et al., 2020) in Pytorch (Paszke et al., 2019). We train all models with batch size 16, learning rate $2e-5$, no weight decay, AdamW optimizer with epsilon $1e-8$. For all datasets we use maximum sequence length of 128, except for IMDB and AGNEWS that contain longer input texts, where we use 256. To ensure reproducibility and fair comparison between the various methods under evaluation, we run all experiments with the same five seeds that we randomly selected from the range $[1, 9999]$.

A.4 Baselines

Acquisition functions We compare ENTROPYwith four baseline acquisition functions. The first is the standard AL baseline, **RANDOM**, which applies uniform sampling and selects k data points from $\mathcal{D}_{\textbf{pool}}$ at each iteration. The second is **BADGE** (Ash et al., 2020), an acquisition function that aims to combine diversity and uncertainty sampling. The algorithm computes *gradient embeddings* g_x for every candidate data point x in $\mathcal{D}_{\textbf{pool}}$ and then uses clustering to select a batch. Each g_x is computed as the gradient of the cross-entropy loss with respect to the parameters of the model's last layer. We also compare against a recently introduced cold-start acquisition function called **ALPS** (Yuan et al., 2020). ALPS acquisition uses the masked language model (MLM) loss of BERT as a proxy for model uncertainty in the downstream classification task. Specifically, aiming to leverage both uncertainty and diversity, ALPS forms a *surprisal embedding* s_x for each x, by passing the unmasked input x through the BERT MLM head to compute the cross-entropy loss for a random 15% subsample of tokens against the target labels. ALPS clusters these embeddings to sample k sentences for each AL iteration. Last, following Yuan et al. (2020), we use **BERTKM** as a diversity baseline, where the l_2 normalized BERT output embeddings are used for clustering.

Models & Fine-tuning Methods We evaluate two variants of the pretrained language model; the original **BERT** model, used in Yuan et al. (2020) and Ein-Dor et al. (2020)[6], and our adapted model **BERT-TAPT** (§2), and two fine-tuning methods;

[6] Ein-Dor et al. (2020) evaluate various acquisition functions, including entropy with MC dropout, and use BERT with the standard fine-tuning approach (SFT).

our proposed fine-tuning approach **FT+** (§2) and standard BERT fine-tuning **SFT**.

MODEL	TREC-6	DBPEDIA	IMDB	SST-2	AGNEWS
VALIDATION SET					
BERT	94.4	99.1	90.7	93.7	94.4
BERT-TAPT	95.2	99.2	91.9	94.3	94.5
TEST SET					
BERT	80.6	99.2	91.0	90.6	94.0
BERT-TAPT	77.2	99.2	91.9	90.8	94.2

Table 2: Accuracy with 100% of data over five runs (different random seeds).

B Appendix: Analysis

B.1 Task-Adaptive Pretraining (TAPT) & Full-Dataset Performance

As discussed in §2 and §4, we continue training the BERT-BASE (Devlin et al., 2019) pretrained masked language model using the available data $\mathcal{D}_{\textbf{pool}}$. We explored various learning rates between $1e-4$ and $1e-5$ and found the latter to produce the lowest validation loss. We trained each model (one for each dataset) for up to 100K optimization steps, we evaluated on $\mathcal{D}_{\textbf{val}}$ every 500 steps and saved the checkpoint with the lowest validation loss. We used the resulting model in our (BERT-TAPT) experiments. We plot the learning curves of masked language modeling task (TAPT) for three datasets and all considered learning rates in Figure 5. We notice that a smaller learning rate facilitates the training of the MLM.

In Table 2 we provide the validation and test accuracy of BERT and BERT-TAPT for all datasets. We present the mean across runs with three random seeds. For fine-tuning the models, we used the proposed approach FT+ (§2).

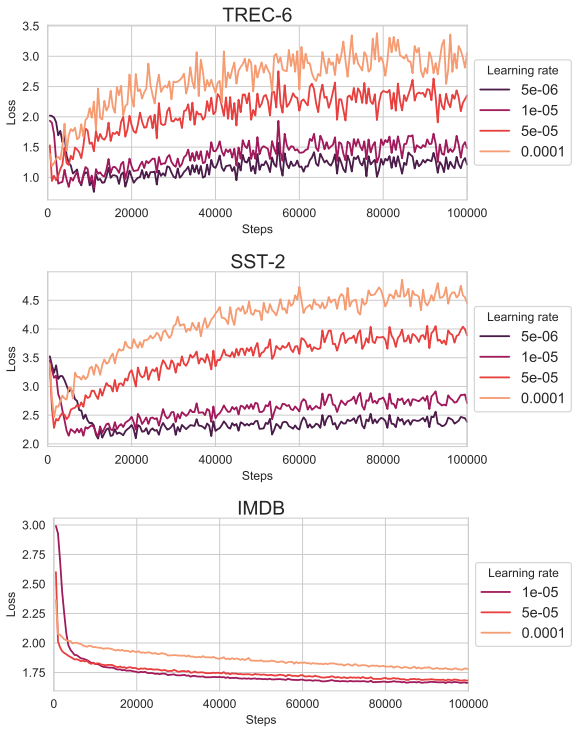


Figure 5: Learning curves of TAPT for various learning rates.

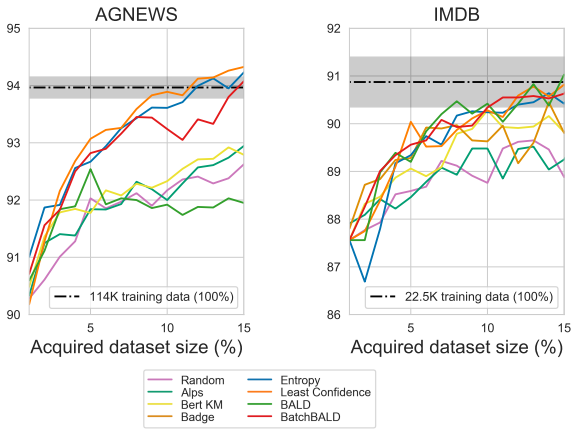


Figure 6: Comparison of acquisition functions using TAPT and FT+ in training BERT.

B.2 Performance of Acquisition Functions

In our BERT-TAPT-FT+ experiments so far, we showed results with ENTROPY. We have also experimented with various uncertainty-based acquisition functions. Specifically, four uncertainty-based acquisition functions are used in our work: LEAST CONFIDENCE, ENTROPY, BALD and BATCHBALD. LEAST CONFIDENCE (Lewis and Gale, 1994) sorts $\mathcal{D}_{pool}$ by the probability of *not* predicting the most confident class, in descending order, ENTROPY (Shannon, 1948) selects samples that maximize the predictive entropy, and BALD (Houlsby et al., 2011), short for Bayesian Active Learning by Disagreement, chooses data points that maximize the mutual information between predictions and model's posterior probabilities. BATCHBALD (Kirsch et al., 2019) is a recently introduced extension of BALD that *jointly* scores points by estimating the mutual information between multiple data points and the model parameters. This iterative algorithm aims to find *batches* of informative data points, in contrast to BALD that chooses points that are informative individually. Note that LEAST CONFIDENCE, ENTROPY and BALD have been used in AL for NLP by Siddhant and Lipton (2018). To the best of our

	TREC-6	SST-2	IMDB	DBPEDIA	AGNEWS
RANDOM	0/0	0/0	0/0	0/0	0/0
ALPS	0/57	0/478	0/206	0/134	0/634
BADGE	0/63	0/23110	0/1059	0/192	-
BERTKM	0/47	0/2297	0/324	0/137	0/3651
ENTROPY	81/0	989/0	557/0	264/0	2911/0
LEAST CONFIDENCE	69/0	865/0	522/0	256/0	2607/0
BALD	69/0	797/0	524/0	256/0	2589/0
BATCHBALD	69/21	841/1141	450/104	256/482	2844/5611

Table 3: Runtimes (in seconds) for all datasets. In each cell of the table we present a tuple i/s where i is the *inference time* and s the *selection time*. *Inference time* is the time for the model to perform a forward pass for all the unlabeled data in $\mathcal{D}_{\textbf{pool}}$ and *selection time* is the time that each acquisition function requires to rank all candidate data points and select k for annotation (for a single iteration). Since we cannot report the runtimes for every model in the AL pipeline (at each iteration the size of $\mathcal{D}_{\textbf{pool}}$ changes), we provide the median.

knowledge, BATCHBALD is evaluated for the first time in the NLP domain.

Instead of using the output softmax probabilities for each class, we use a probabilistic formulation of deep neural networks in order to acquire better calibrated scores. Monte Carlo (MC) dropout (Gal and Ghahramani, 2016) is a simple yet effective method for performing approximate variational inference, based on dropout (Srivastava et al., 2014). Gal and Ghahramani (2016) prove that by simply performing *dropout during the forward pass in making predictions*, the output is equivalent to the prediction when the parameters are sampled from a variational distribution of the true posterior. Therefore, dropout during inference results into obtaining predictions from different parts of the network. Our BERT-based $\mathcal{M}_i$ model uses dropout layers during training for regularization. We apply MC dropout by simply activating them during test time and we perform multiple stochastic forward passes. Formally, we do N passes of every $x \in \mathcal{D}_{\textbf{pool}}$ through $\mathcal{M}_i(x; W_i)$ to acquire N different output probability distributions for each x. MC dropout for AL has been previously used in the literature (Gal et al., 2017; Shen et al., 2017; Siddhant and Lipton, 2018; Lowell and Lipton, 2019; Ein-Dor et al., 2020; Shelmanov et al., 2021).

Our findings show that all functions provide similar performance, except for BALD that slightly underperforms. This makes our approach agnostic to the selected uncertainty-based acquisition method. We also evaluate our proposed methods with our baseline acquisition functions, i.e. RANDOM, ALPS, BERTKM and BADGE, since our training strategy is orthogonal to the acquisition strategy. We compare all acquisition functions with BERT-TAPT-FT+ for AGNEWS and IMDB in Figure 6. We observe that in general uncertainty-based acquisition performs better compared to diversity, while all acquisition strategies have benefited from our training strategy (TAPT and FT+).

B.3 Efficiency of Acquisition Functions

In this section we discuss the efficiency of the eight acquisition functions considered in this work; RANDOM, ALPS, BADGE, BERTKM, ENTROPY, LEAST CONFIDENCE, BALD and BATCHBALD.

In Table 3 we provide the runtimes for all acquisition functions and datasets. Each AL experiments consists of multiple iterations and (therefore multiple models), each with a different training dataset $\mathcal{D}_{\textbf{lab}}$ and pool of unlabeled data $\mathcal{D}_{\textbf{pool}}$. In order to evaluate how computationally heavy is each method, we provide the *median* of all the models in one AL experiment. We calculate the runtime of two types of functionalities. The first is the *inference time* and stands for the forward pass of each $x \in \mathcal{D}_{\textbf{pool}}$ to acquire confidence scores for uncertainty sampling. RANDOM, ALPS, BADGE and BERTKM do not require this step so it is only applied of uncertainty-based acquisition where acquiring uncertainty estimates with MC dropout is needed. The second functionality is *selection time* and measures how much time each acquisition function requires to rank and select the k data points from $\mathcal{D}_{\textbf{pool}}$ to be labeled in the next step of the AL pipeline. RANDOM, ENTROPY, LEAST CONFIDENCE and BALD perform simple equations to rank the data points and therefore so do not require selection time. On the other hand, ALPS, BADGE,

BERTKM and BATCHBALD perform iterative algorithms that increase selection time. From all acquisition functions ALPS and BERTKM are faster because they do not require the inference step of all the unlabeled data to the model. ENTROPY, LEAST CONFIDENCE and BALD require the same time for selecting data, which is equivalent for the time needed to perform one forward pass of the entire $\mathcal{D}_{\textbf{pool}}$. Finally BADGE and BATCHBALD are the most computationally heavy approaches, since both algorithms require multiple computations for the *selection time*. RANDOM has a total runtime of zero seconds, as expected.

2.3 Impact

According to Google Scholar, the paper has received 41 citations as of May 2024. It was featured in numerous surveys (e.g. Tsvigun et al., 2022; Zhang et al., 2022f; Rauch et al., 2023; Mehlin et al., 2023; Wang et al., 2023; Wan et al., 2023; Nachtegael et al., 2023; Li et al., 2024; Tamkin, 2023; Zhang, 2023; Rainforth et al., 2024) and has inspired follow-up work (Steegh and Sileno, 2023; Shi and Lipani, 2023; Shi et al., 2023).

2.4 Discussion

Apart from considerably improving active learning with BERT models, the more general takeaways of the paper are twofold: (i) adapting a pretrained language model that will be used for AL to the domain of the task is crucial and efficient, if such unlabeled data is available beforehand, and (ii) the training strategy can be more influential in the AL success than the acquisition strategy (i.e., if a model is high performing – such as through our proposed adaptation methodology – several acquisition functions paired with it provide similar results).

Naturally the field of NLP has advanced significantly during the years that this PhD thesis has been conducted, and BERT-like models are often replaced by large language models (LLMs) of million or billion of parameters that are based on the Transformer architecture. Still, our key findings in this paper are general enough to aid practitioners in building effective AL methods for their work. Using a model that is highly capable in the domain of the task at hand is most likely always desired, as a very large pretrained model still has limited capacity to perform perfectly in all scenarios. Also, even though the exact fine-tuning technique we proposed will most likely be deprecated now, our idea that the varying dataset size of the acqtively acquired training dataset during the AL iterations will always be useful to take into account when developing the AL training methodology.

Chapter 3

Publication II: Constrastive Active Learning

The main contribution of this chapter is the paper *Active Learning by Acquiring Contrastive Examples*, which was published at the *Empirical Methods in Natural Language Processing* conference in November 2021. We first outline the motivation for this work (Section 3.1), followed by the paper itself (Section 3.2), the impact that it has had so far (Section 3.3), and discussion (Section 3.4).

3.1 Introduction

Following our contribution to the training stage of the active learning (AL) loop (Chapter 2), we turned our attention to the data acquisition stage. Selecting the appropriate data for human annotation is arguably the most crucial aspect of AL. We examined the most prevalent uncertainty-based and diversity-based AL acquisition methods, critically analyzing their limitations. Existing methods often either select hard examples that are impossible for the model to learn (e.g., collective outliers) or data points that are diverse but too easy for the model to learn, thus not significantly enhancing data efficiency. To address this gap in the literature, we proposed Contrastive Active Learning (CAL). CAL is an algorithm that selects examples near the model's decision boundary, using the feature space of the hidden representations to determine which data points will most effectively enhance the training set.

Section 2 of our paper below describes in detail how we formulated the algorithm for CAL. The idea is to find, for each example in the unlabeled pool x_p, a neighbourhood of k data points in the labeled dataset that the model produces the most diverging predictions. We would expect that the k nearest neighbours $x_{l_i^k}$ of x_p to belong to

the same class as x_p. We measure the model's predicted probability for $p(y|x_l)_i^k$ and compute the KL divergence with $p(y|x_p)$. We compute a score as the average KL divergence of the neighborhood for each x_p in the pool. In the end, we choose to acquire the data points whose average KL score was the highest, as these examples diverge the most from their neighbours in the training set, being therefore close to the model's decision boundary (i.e. contrastive examples).

Through empirical experimentation we found that CAL the top performing (in some cases along with the entropy acquisition function) algorithm between 6 methods, yielding high AL results in several text classification tasks. We conducted a thorough analysis of the algorithm (Sections 5 and 6 of the paper) and showed that CAL chooses uncertain yet representative data.

3.2 The Paper

Author Contributions

The paper is co-authored by myself, Giorgos Vernikos, Loïc Barrault and Nikolaos Aletras. Nikolaos Aletras seeded the idea of exloring a more effective acquisition function for active learning, co-supervised the project, offered suggestions, and helped revise the final version. Loïc Barrault co-supervised the project, attended all meetings with Nikolaos Aletras and myself for the project, helped shape the idea for the algorithm and proofread the paper. As the lead author, I developed the algorithm for CAL, performed the experiments, and wrote the paper. Giorgos Vernikos participated in many discussions, helped brainstorming the algorithm for CAL and gave pointers for analysing the method and the results.

Active Learning by Acquiring Contrastive Examples

Katerina Margatina[†] **Giorgos Vernikos**[‡*] **Loïc Barrault**[†] **Nikolaos Aletras**[†]
[†]University of Sheffield [‡]EPFL [*]HEIG-VD
{k.margatina, l.barrault, n.aletras}@sheffield.ac.uk
georgios.vernikos@epfl.ch

Abstract

Common acquisition functions for active learning use either uncertainty or diversity sampling, aiming to select difficult and diverse data points from the pool of unlabeled data, respectively. In this work, leveraging the best of both worlds, we propose an acquisition function that opts for selecting *contrastive examples*, i.e. data points that are similar in the model feature space and yet the model outputs maximally different predictive likelihoods. We compare our approach, CAL (Contrastive Active Learning), with a diverse set of acquisition functions in four natural language understanding tasks and seven datasets. Our experiments show that CAL performs consistently better or equal than the best performing baseline across all tasks, on both in-domain and out-of-domain data. We also conduct an extensive ablation study of our method and we further analyze all actively acquired datasets showing that CAL achieves a better trade-off between uncertainty and diversity compared to other strategies.

1 Introduction

Active learning (AL) is a machine learning paradigm for efficiently acquiring data for annotation from a (typically large) pool of unlabeled data (Lewis and Catlett, 1994; Cohn et al., 1996; Settles, 2009). Its goal is to concentrate the human labeling effort on the most informative data points that will benefit model performance the most and thus reducing data annotation cost.

The most widely used approaches to acquiring data for AL are based on uncertainty and diversity, often described as the "two faces of AL" (Dasgupta, 2011). While uncertainty-based methods leverage the model predictive confidence to select difficult examples for annotation (Lewis and Gale, 1994; Cohn et al., 1996), diversity sampling exploits heterogeneity in the feature space by typically performing clustering (Brinker, 2003; Bodó et al., 2011). Still, both approaches have core limitations that may lead to acquiring redundant data points. Algorithms based on uncertainty may end up choosing uncertain yet uninformative repetitive data, while diversity-based methods may tend to select diverse yet easy examples for the model (Roy and McCallum, 2001). The two approaches are orthogonal to each other, since uncertainty sampling is usually based on the model's output, while diversity exploits information from the input (i.e. feature) space. Hybrid data acquisition functions that combine uncertainty and diversity sampling have also been proposed (Shen et al., 2004; Zhu et al., 2008; Ducoffe and Precioso, 2018; Ash et al., 2020; Yuan et al., 2020; Ru et al., 2020).

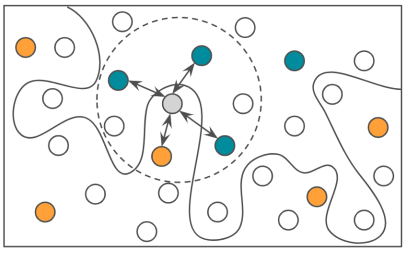


Figure 1: Illustrative example of our proposed method CAL. The solid line (model decision boundary) separates data points from two different classes (blue and orange), the coloured data points represent the labeled data and the rest are the unlabeled data of the pool.

In this work, we aim to leverage characteristics from hybrid data acquisition. We hypothesize that data points that are close in the model feature space (i.e. share similar or related vocabulary, or similar model encodings) but the model produces different predictive likelihoods, should be good candidates for data acquisition. We define such examples as *contrastive* (see example in Figure 1). For that purpose, we propose a new acquisition function that searches for contrastive examples in the pool of unlabeled data. Specifically, our method, Contrastive Active Learning (CAL) *selects unlabeled*

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 650–663
November 7–11, 2021.

data points from the pool, whose predictive likelihoods diverge the most from their neighbors in the training set. This way, CAL shares similarities with diversity sampling, but instead of performing clustering it uses the feature space to create neighborhoods. CAL also leverages uncertainty, by using predictive likelihoods to rank the unlabeled data.

We evaluate our approach in seven datasets from four tasks including sentiment analysis, topic classification, natural language inference and paraphrase detection. We compare CAL against a full suite of baseline acquisition functions that are based on uncertainty, diversity or both. We also examine robustness by evaluating on out-of-domain data, apart from in-domain held-out sets. Our contributions are the following:

1. We propose CAL, a new acquisition function for active learning that acquires contrastive examples from the pool of unlabeled data (§2);
2. We show that CAL performs consistently better or equal compared to all baselines in all tasks when evaluated on in-domain and out-of-domain settings (§4);
3. We conduct a thorough analysis of our method showing that CAL achieves a better trade-off between diversity and uncertainty compared to the baselines (§6).

We release our code online [1].

2 Contrastive Active Learning

In this section we present in detail our proposed method, CAL: Contrastive Active Learning. First, we provide a definition for contrastive examples and how they are related to finding data points that are close to the decision boundary of the model (§2.1). We next describe an active learning loop using our proposed acquisition function (§2.2).

2.1 Contrastive Examples

In the context of active learning, we aim to formulate an acquisition function that selects contrastive examples from a pool of unlabeled data for annotation. We draw inspiration from the contrastive learning framework, that leverages the similarity between data points to push those from the same class closer together and examples from different classes further apart during training (Mikolov et al., 2013; Sohn, 2016; van den Oord et al., 2019; Chen et al., 2020; Gunel et al., 2021).

In this work, we define as contrastive examples two data points if their model encodings are similar, but their model predictions are very different (maximally disagreeing predictive likelihoods).

Formally, data points x_i and x_j should first satisfy a similarity criterion:

$$d\big(\Phi(x_i), \Phi(x_j)\big) < \epsilon \tag{1}$$

where $\Phi(.) \in \mathbb{R}^{d'}$ is an encoder that maps x_i, x_j in a shared feature space, $d(.)$ is a distance metric and ϵ is a small distance value.

A second criterion, based on model uncertainty, is to evaluate that the predictive probability distributions of the model $p(y|x_i)$ and $p(y|x_j)$ for the inputs x_i and x_j should maximally diverge:

$$\mathrm{KL}\big(p(y|x_i)||p(y|x_j)\big) \rightarrow \infty \tag{2}$$

where KL is the Kullback-Leibler divergence between two probability distributions [2].

For example, in a binary classification problem, given a reference example x_1 with output probability distribution $(0.8, 0.2)$ [3] and similar candidate examples x_2 with $(0.7, 0.3)$ and x_3 with $(0.6, 0.4)$, we would consider as contrastive examples the pair (x_1, x_3). However, if another example x_4 (similar to x_1 in the model feature space) had a probability distribution $(0.4, 0.6)$, then the most contrastive pair would be (x_1, x_4).

Figure 1 provides an illustration of contrastive examples for a binary classification case. All data points inside the circle (dotted line) are similar in the model feature space, satisfying Eq. 1. Intuitively, if the divergence of the output probabilities of the model for the gray and blue shaded data points is high, then Eq. 2 should also hold and we should consider them as contrastive.

From a different perspective, data points with similar model encodings (Eq. 1) and dissimilar model outputs (Eq. 2), should be close to the model's decision boundary (Figure 1). Hence, we hypothesize that our proposed approach to select

[1] https://github.com/mourga/contrastive-active-learning

[2] KL divergence is not a symmetric metric, $\mathrm{KL}(P||Q) = \sum_x P(x)\log\big(\frac{P(x)}{Q(x)}\big)$. We use as input Q the output probability distribution of an unlabeled example from the pool and as target P the output probability distribution of an example from the train set (See §2.2 and algorithm 1).

[3] A predictive distribution $(0.8, 0.2)$ here denotes that the model is 80% confident that x_1 belongs to the first class and 20% to the second.

Algorithm 1 Single iteration of CAL

Input: labeled data $\mathcal{D}_{\textbf{lab}}$, unlabeled data $\mathcal{D}_{\textbf{pool}}$, acquisition size b, model $\mathcal{M}$, number of neighbours k, model representation (encoding) function $\Phi(.)$

1 **for** x_p *in* $\mathcal{D}_{\textbf{pool}}$ **do**
2 $\quad \{(x_l^{(i)}, y_l^{(i)})\}, i = 1, ..., k \leftarrow \text{KNN}(\Phi(x_p), \Phi(\mathcal{D}_{\textbf{lab}}), k)$ ▷ find neighbours in $\mathcal{D}_{\textbf{lab}}$
3 $\quad p(y|x_l^{(i)}) \leftarrow \mathcal{M}(x_l^{(i)}), i = 1, ..., k$ ▷ compute probabilities
4 $\quad p(y|x_p) \leftarrow \mathcal{M}(x_p)$
5 $\quad \text{KL}(p(y|x_l^{(i)})||p(y|x_p)), i = 1, ..., k$ ▷ compute divergence
6 $\quad s_{x_p} = \frac{1}{k} \sum_{i=1}^{k} \text{KL}(p(y|x_l^{(i)})||p(y|x_p))$ ▷ score
7 **end**
8 $Q = \underset{x_p \in \mathcal{D}_{\textbf{pool}}}{\text{argmax}} s_{x_p}, |Q| = b$ ▷ select batch

Output: Q

contrastive examples is related to acquiring difficult examples near the decision boundary of the model. Under this formulation, CAL does not guarantee that the contrastive examples lie near the model's decision boundary, because our definition is not strict. In order to ensure that a pair of contrastive examples lie on the boundary, the second criterion should require that the model classifies the two examples in different classes (i.e. different predictions). However, calculating the distance between an example and the model decision boundary is intractable and approximations that use adversarial examples are computationally expensive (Ducoffe and Precioso, 2018).

2.2 Active Learning Loop

Assuming a multi-class classification problem with C classes, labeled data for training $\mathcal{D}_{\textbf{lab}}$ and a pool of unlabeled data $\mathcal{D}_{\textbf{pool}}$, we perform AL for T iterations. At each iteration, we train a model on $\mathcal{D}_{\textbf{lab}}$ and then use our proposed acquisition function, CAL (Algorithm 1), to acquire a batch Q consisting of b examples from $\mathcal{D}_{\textbf{pool}}$. The acquired examples are then labeled[4], they are removed from the pool $\mathcal{D}_{\textbf{pool}}$ and added to the labeled dataset $\mathcal{D}_{\textbf{lab}}$, which will serve as the training set for training a model in the next AL iteration. In our experiments, we use a pretrained BERT model $\mathcal{M}$ (Devlin et al., 2019), which we fine-tune at each AL iteration using the current $\mathcal{D}_{\textbf{lab}}$. We begin the AL loop by training a model $\mathcal{M}$ using an initial labeled dataset $\mathcal{D}_{\textbf{lab}}$ [5].

Find Nearest Neighbors for Unlabeled Candidates The first step of our contrastive acquisition function (cf. line 2) is to find examples that are similar in the model feature space (Eq. 1). Specifically, we use the `[CLS]` token embedding of BERT as our encoder $\Phi(.)$ to represent all data points in $\mathcal{D}_{\textbf{lab}}$ and $\mathcal{D}_{\textbf{pool}}$. We use a K-Nearest-Neighbors (KNN) implementation using the labeled data $\mathcal{D}_{\textbf{lab}}$, in order to query similar examples $x_l \in \mathcal{D}_{\textbf{lab}}$ for each candidate $x_p \in \mathcal{D}_{\textbf{pool}}$. Our distance metric $d(.)$ is Euclidean distance. To find the most similar data points in $\mathcal{D}_{\textbf{lab}}$ for each x_p, we select the top k instead of selecting a predefined threshold ϵ (Eq. 1) [6]. This way, we create a neighborhood $N_{x_p} = \{x_p, x_l^{(1)}, \ldots, x_l^{(k)}\}$ that consists of the unlabeled data point x_p and its k closest examples x_l in $\mathcal{D}_{\textbf{lab}}$ (Figure 1).

Compute Contrastive Score between Unlabeled Candidates and Neighbors In the second step, we compute the divergence in the model predictive probabilities for the members of the neighborhood (Eq. 2). Using the current trained model $\mathcal{M}$ to obtain the output probabilities for all data points in N_{x_p} (cf. lines 3-4), we then compute the Kullback–Leibler divergence (KL) between the output probabilities of x_p and all $x_l \in N_{x_p}$ (cf. line 5). To obtain a score s_{x_p} for a candidate x_p, we take the average of all divergence scores (cf. line 6).

Rank Unlabeled Candidates and Select Batch We apply these steps to all candidate examples $x_p \in \mathcal{D}_{\textbf{pool}}$ and obtain a score s_{x_p} for each. With

[4] We simulate AL, so we already have the labels of the examples of $\mathcal{D}_{\textbf{pool}}$ (but still treat it as an unlabeled dataset).

[5] We acquire the first examples that form the initial training set $\mathcal{D}_{\textbf{lab}}$ by applying random stratified sampling (i.e. keeping the initial label distribution).

[6] We leave further modifications of our scoring function as future work. One approach would be to add the average distance from the neighbors (cf. line 6) in order to alleviate the possible problem of selecting outliers.

DATASET	TASK	DOMAIN	OOD DATASET	TRAIN	VAL	TEST	CLASSES
IMDB	Sentiment Analysis	Movie Reviews	SST-2	22.5K	2.5K	25K	2
SST-2	Sentiment Analysis	Movie Reviews	IMDB	60.6K	6.7K	871	2
AGNEWS	Topic Classification	News	-	114K	6K	7.6K	4
DBPEDIA	Topic Classification	News	-	20K	2K	70K	14
PUBMED	Topic Classification	Medical	-	180K	30.2K	30.1K	5
QNLI	Natural Language Inference	Wikipedia	-	99.5K	5.2K	5.5K	2
QQP	Paraphrase Detection	Social QA Questions	TWITTERPPDB	327K	36.4K	80.8K	2

Table 1: Dataset statistics.

our scoring function we define as contrastive examples the unlabeled data x_p that have the highest score s_{x_p}. A high s_{x_p} score indicates that the unlabeled data point x_p has a high divergence in model predicted probabilities compared to its neighbors in the training set (Eq. 1, 2), suggesting that it may lie near the model's decision boundary. To this end, our acquisition function selects the top b examples from the pool that have the highest score s_{x_p} (cf. line 8), that form the acquired batch Q.

3 Experimental Setup

3.1 Tasks & Datasets

We conduct experiments on sentiment analysis, topic classification, natural language inference and paraphrase detection tasks. We provide details for the datasets in Table 1. We follow Yuan et al. (2020) and use IMDB (Maas et al., 2011), SST-2 (Socher et al., 2013), PUBMED (Dernoncourt and Lee, 2017) and AGNEWS from Zhang et al. (2015) where we also acquired DBPEDIA. We experiment with tasks requiring pairs of input sequences, using QQP and QNLI from GLUE (Wang et al., 2019). To evaluate robustness on out-of-distribution (OOD) data, we follow Hendrycks et al. (2020) and use SST-2 as OOD dataset for IMDB and vice versa. We finally use TWITTERPPDB (Lan et al., 2017) as OOD data for QQP as in Desai and Durrett (2020).

3.2 Baselines

We compare CAL against five baseline acquisition functions. The first method, ENTROPY is the most commonly used uncertainty-based baseline that acquires data points for which the model has the highest predictive entropy. As a diversity-based baseline, following Yuan et al. (2020), we use BERTKM that applies k-means clustering using the l_2 normalized BERT output embeddings of the fine-tuned model to select b data points. We compare against BADGE (Ash et al., 2020), an acquisition function that aims to combine diversity and uncertainty sampling, by computing *gradient embeddings* g_x for every candidate data point x in $\mathcal{D}_{\textbf{pool}}$ and then using clustering to select a batch. Each g_x is computed as the gradient of the cross-entropy loss with respect to the parameters of the model's last layer, aiming to be the component that incorporates uncertainty in the acquisition function [7]. We also evaluate a recently introduced cold-start acquisition function called ALPS (Yuan et al., 2020) that uses the masked language model (MLM) loss of BERT as a proxy for model uncertainty in the downstream classification task. Specifically, aiming to leverage both uncertainty and diversity, ALPS forms a *surprisal embedding* s_x for each x, by passing the unmasked input x through the BERT MLM head to compute the cross-entropy loss for a random 15% subsample of tokens against the target labels. ALPS clusters these embeddings to sample b sentences for each AL iteration. Lastly, we include RANDOM, that samples data from the pool from a uniform distribution.

3.3 Implementation Details

We use BERT-BASE (Devlin et al., 2019) adding a task-specific classification layer using the implementation from the HuggingFace library (Wolf et al., 2020). We evaluate the model 5 times per epoch on the development set following Dodge et al. (2020) and keep the one with the lowest validation loss. We use the standard splits provided for all datasets, if available, otherwise we randomly sample a validation set from the training set. We test all models on a held-out test set. We repeat all experiments with five different random seeds resulting into different initializations of the parameters of the model's extra task-specific output feedfor-

[7] We note that BERTKM and BADGE are computationally heavy approaches that require clustering of vectors with high dimensionality, while their complexity grows exponentially with the acquisition size. We thus do not apply them to the datasets that have a large $\mathcal{D}_{\textbf{pool}}$. More details can be found in the Appendix A.2

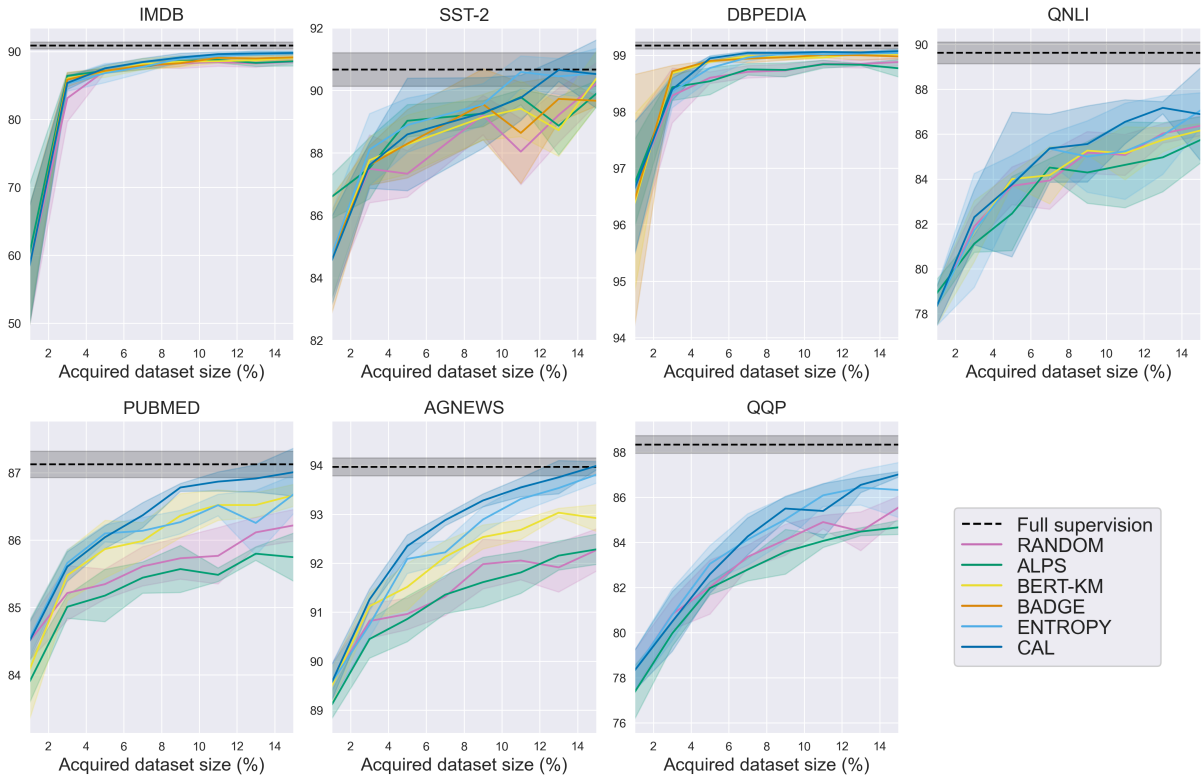


Figure 2: In-domain (ID) test accuracy during AL iterations for different acquisition functions.

ward layer and the initial $\mathcal{D}_{\textbf{lab}}$. For all datasets we use as budget the 15% of $\mathcal{D}_{\textbf{pool}}$, initial training set 1% and acquisition size $b = 2\%$. Each experiment is run on a single Nvidia Tesla V100 GPU. More details are provided in the Appendix A.1.

4 Results

4.1 In-domain Performance

We present results for in-domain test accuracy across all datasets and acquisition functions in Figure 2. We observe that CAL is consistently the top performing method especially in DBPEDIA, PUBMED and AGNEWS datasets.

CAL performs slightly better than ENTROPY in IMDB, QNLI and QQP, while in SST-2 most methods yield similar results. ENTROPY is the second best acquisition function overall, consistently performing better than diversity-based or hybrid baselines. This corroborates recent findings from Desai and Durrett (2020) that BERT is sufficiently calibrated (i.e. produces good uncertainty estimates), making it a tough baseline to beat in AL.

BERTKM is a competitive baseline (e.g. SST-2, QNLI) but always underperforms compared to CAL and ENTROPY, suggesting that uncertainty is the most important signal in the data selection process. An interesting future direction would be to investigate in depth whether and which (i.e. which layer) representations of the current (pretrained language models) works best with similarity search algorithms and clustering.

Similarly, we can see that BADGE, despite using both uncertainty and diversity, also achieves low performance, indicating that clustering the constructed gradient embeddings does not benefit data acquisition. Finally, we observe that ALPS generally underperforms and is close to RANDOM. We can conclude that this heterogeneous approach to uncertainty, i.e. using the pretrained language model as proxy for the downstream task, is beneficial only in the first few iterations, as shown in Yuan et al. (2020).

Surprisingly, we observe that for the SST-2 dataset ALPS performs similarly with the highest performing acquisition functions, CAL and ENTROPY. We hypothesize that due to the informal textual style of the reviews of SST-2 (noisy social media data), the pretrained BERT model can be used as a signal to query linguistically hard examples, that benefit the downstream sentiment analysis task. This is an interesting finding and a future research direction would be to investigate the correlation between the difficulty of an example in a

TRAIN (ID)	SST-2	IMDB	QQP
TEST (OOD)	IMDB	SST-2	TWITTERPPDB
RANDOM	76.28 ± 0.72	82.50 ± 3.61	85.86 ± 0.48
BERTKM	75.99 ± 1.01	84.98 ± 1.22	-
ENTROPY	75.38 ± 2.04	**85.54** ± 2.52	85.06 ± 1.96
ALPS	77.06 ± 0.78	83.65 ± 3.17	84.79 ± 0.49
BADGE	76.41 ± 0.92	85.19 ± 3.01	-
CAL	**79.00** ± 1.39	84.96 ± 2.36	**86.20** ± 0.22

Table 2: Out-of-domain (OOD) accuracy of models trained with the actively acquired datasets created with different AL acquisition strategies.

downstream task with its perplexity (loss) of the pretrained language model.

4.2 Out-of-domain Performance

We also evaluate the out-of-domain (OOD) robustness of the models trained with the actively acquired datasets of the last iteration (i.e. 15% of $\mathcal{D}_{\textbf{pool}}$ or 100% of the AL budget) using different acquisition strategies. We present the OOD results for SST-2, IMDB and QQP in Table 2. When we test the models trained with SST-2 on IMDB (first column) we observe that CAL achieves the highest performance compared to the other methods by a large margin, indicating that acquiring contrastive examples can improve OOD generalization. In the opposite scenario (second column), we find that the highest accuracy is obtained with ENTROPY. However, similarly to the ID results for SST-2 (Figure 2), all models trained on different subsets of the IMDB dataset result in comparable performance when tested on the small SST-2 test set (the mean accuracies lie inside the standard deviations across models). We hypothesize that this is because SST-2 is not a challenging OOD dataset for the different IMDB models. This is also evident by the high OOD accuracy, 85% on average, which is close to the 91% SST-2 ID accuracy of the full model (i.e. trained on 100% of the ID data). Finally, we observe that CAL obtains the highest OOD accuracy for QQP compared to RANDOM, ENTROPY and ALPS. Overall, our empirical results show that the models trained on the actively acquired dataset with CAL obtain consistently similar or better performance than all other approaches when tested on OOD data.

5 Ablation Study

We conduct an extensive ablation study in order to provide insights for the behavior of every component of CAL. We present all AL experiments on the AGNEWS dataset in Figure 3.

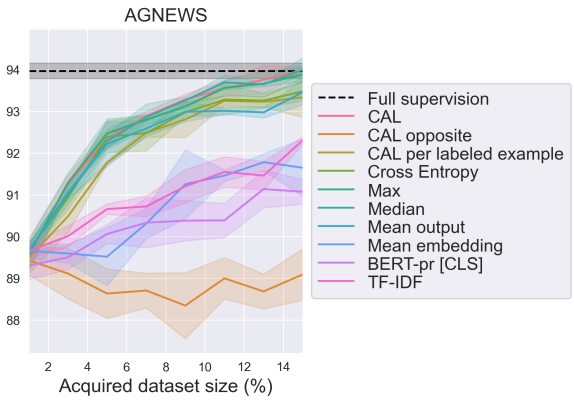


Figure 3: In-domain (ID) test accuracy with different variants of CAL (ablation).

Decision Boundary We first aim to evaluate our hypothesis that CAL acquires difficult examples that lie close to the model's decision boundary. Specifically, to validate that the ranking of the constructed neighborhoods is meaningful, we run an experiment where we acquire candidate examples that have the *minimum* divergence from their neighbors opposite to CAL (i.e. we replace $\text{argmax}(.)$ with $\text{argmin}(.)$ in line 8 of Algorithm 1). We observe (Fig. 3 - `CAL opposite`) that even after acquiring 15% of unlabeled data, the performance remains unchanged compared to the initial model (of the first iteration), even degrades. In effect, this finding denotes that CAL does select informative data points.

Neighborhood Next, we experiment with changing the way we construct the neighborhoods, aiming to improve computational efficiency. We thus modify our algorithm to create a neighborhood for each *labeled* example (instead of unlabeled).[8]. This way we compute a divergence score only for the neighbors of the training data points. However, we find this approach to slightly underperform (Fig. 3 - `CAL per labeled example`), possibly because only a small fraction of the pool is considered and thus the uncertainty of all the unlabeled data points is not taken into account.

[8] In this experiment, we essentially change the *for-loop* of Algorithm 1 (cf. line 1-7) to iterate for each x_l in $\mathcal{D}_{\textbf{lab}}$ (instead of each x_p in $\mathcal{D}_{\textbf{pool}}$) and similarly find the k nearest neighbors of each labeled example in the pool ($\text{KNN}(x_l, \mathcal{D}_{\textbf{pool}}, k)$) As for the scoring (cf. line 6), if an unlabeled example was not picked (i.e. was not a neighbor to a labeled example), its score is zero. If it was picked multiple times we average its scores. We finally acquire the top b unlabeled data with the highest scores. This formulation is more computationally efficient since usually $|\mathcal{D}_{\textbf{lab}}| << |\mathcal{D}_{\textbf{pool}}|$.

Scoring function We also experiment with several approaches for constructing our scoring function (cf. line 6 in Algorithm 1). Instead of computing the KL divergence between the predicted probabilities of each candidate example and its labeled neighbors (cf. line 5), we used cross entropy between the output probability distribution and the gold labels of the labeled data. The intuition is to evaluate whether information of the actual label is more useful than the model's predictive probability distribution. We observe this scoring function to result in a slight drop in performance (Fig. 3 - `Cross Entropy`). We also experimented with various pooling operations to aggregate the KL divergence scores for each candidate data point. We found maximum and median (Fig. 3 - `Max/Median`) to perform similarly with the average (Fig. 3 - CAL), which is the pooling operation we decided to keep in our proposed algorithm.

Feature Space Since our approach is related to to acquiring data near the model's decision boundary, this effectively translates into using the `[CLS]` output embedding of BERT. Still, we opted to cover several possible alternatives to the representations, i.e. feature space, that can be used to find the neighbors with KNN. We divide our exploration into two categories: intrinsic representations from the current fine-tuned model and extrinsic using different methods. For the first category, we examine representing each example with the mean embedding layer of BERT (Fig. 3 - `Mean embedding`) or the mean output embedding (Fig. 3 - `Mean output`). We find both alternatives to perform worse than using the `[CLS]` token (Fig. 3 - CAL). The motivation for the second category is to evaluate whether acquiring contrastive examples in the *input* feature space, i.e. representing the raw text, is meaningful (Gardner et al., 2020) [9]. We thus examine contextual representations from a pretrained BERT language model (Fig. 3 - `BERT-pr [CLS]`) (not fine-tuned in the task or domain) and non-contextualized TF-IDF vectors (Fig. 3 - `TF-IDF`). We find both approaches, along with `Mean embedding`, to largely underperform compared to our approach that acquires ambiguous data near the model decision boundary.

[9] This can be interpreted as comparing the effectiveness of selecting data near the *model* decision boundary vs. the *task* decision boundary, i.e. data that are similar for the task itself or for the humans (in terms of having the same raw input/vocabulary), but are from different classes.

6 Analysis

Finally, we further investigate CAL and all acquisition functions considered (baselines), in terms of diversity, representativeness and uncertainty. Our aim is to provide insights on what data each method tends to select and what is the uncertainty-diversity trade-off of each approach. Table 3 shows the results of our analysis averaged across datasets. We denote with L the labeled set, U the unlabeled pool and Q an acquired batch of data points from U [10].

6.1 Diversity & Uncertainty Metrics

Diversity in input space (DIV.-I) We first evaluate the diversity of the actively acquired data in the input feature space, i.e. raw text, by measuring the overlap between tokens in the sampled sentences Q and tokens from the rest of the data pool U. Following Yuan et al. (2020), we compute DIV.-I as the Jaccard similarity between the set of tokens from the sampled sentences Q, $\mathcal{V}_Q$, and the set of tokens from the unsampled sentences $\mathcal{U}\backslash\mathcal{Q}$, $\mathcal{V}_{Q'}$, $\mathcal{J}(\mathcal{V}_Q, \mathcal{V}_{Q'}) = \frac{|\mathcal{V}_Q \cap \mathcal{V}_{Q'}|}{|\mathcal{V}_Q \cup \mathcal{V}_{Q'}|}$. A high DIV.-I value indicates high diversity because the sampled and unsampled sentences have many tokens in common.

Diversity in feature space (DIV.-F) We next evaluate diversity in the (model) feature space, using the `[CLS]` representations of a trained BERT model [11]. Following Zhdanov (2019) and Ein-Dor et al. (2020), we compute DIV.-F of a set Q as $\left(\frac{1}{|U|}\sum_{x_i \in U} \min_{x_j \in Q} d(\Phi(x_i), \Phi(x_j))\right)^{-1}$, where $\Phi(x_i)$ denotes the `[CLS]` output token of example x_i obtained by the model which was trained using L, and $d(\Phi(x_i), \Phi(x_j))$ denotes the Euclidean distance between x_i and x_j in the feature space.

Uncertainty (UNC.) To measure uncertainty, we use the model $\mathcal{M}_f$ trained on the entire training dataset (Figure 2 - `Full supervision`). As in Yuan et al. (2020), we use the logits from the fully trained model to estimate the uncertainty of an example, as it is a reliable estimate due to its high performance after training on many examples, while

[10] In the previous sections we used $\mathcal{D}_{\text{lab}}$ and $\mathcal{D}_{\text{pool}}$ to denote the labeled and unlabeled sets and we change the notation here to L and U, respectively, for simplicity.

[11] To enable an appropriate comparison, this analysis is performed after the initial BERT model is trained with the initial training set and each AL strategy has selected examples equal to 2% of the pool (first iteration). Correspondingly, all strategies select examples from the same unlabeled set U while using outputs from the same BERT model.

	DIV.-I	DIV.-F	UNC.	REPR.
RANDOM	0.766	0.356	0.132	1.848
BERTKM	0.717	**0.363**	0.145	2.062
ENTROPY	0.754	0.323	**0.240**	2.442
ALPS	**0.771**	0.360	0.126	2.038
BADGE	0.655	0.339	0.123	2.013
CAL	0.768	0.335	0.231	**2.693**

Table 3: Uncertainty and diversity metrics across acquisition functions, averaged for all datasets.

it offers a fair comparison across all acquisition strategies. First, we compute predictive entropy of an input x when evaluated by model $\mathcal{M}_f$ and then we take the average over all sentences in a sampled batch Q. We use the average predictive entropy to estimate uncertainty of the acquired batch Q for each method $-\frac{1}{|Q|}\sum_{x\in Q}\sum_{c=1}^{C} p(y=c|x)\log p(y=c|x)$. As a sampled batch Q we use the full actively acquired dataset after completing our AL iterations (with 15% of the data).

Representativeness (REPR.) We finally analyze the representativeness of the acquired data as in Ein-Dor et al. (2020). We aim to study whether AL strategies tend to select outlier examples that do not properly represent the overall data distribution. We rely on the KNN-density measure proposed by Zhu et al. (2008), where the density of an example is quantified by one over the average distance between the example and its K most similar examples (i.e., K nearest neighbors) within U, based on the `[CLS]` representations as in DIV.-F. An example with high density degree is less likely to be an outlier. We define the representativeness of a batch Q as one over the average KNN-density of its instances using the Euclidean distance with K=10.

6.2 Discussion

We first observe in Table 3 that ALPS acquires the most diverse data across all approaches. This is intuitive since ALPS is the most linguistically-informed method as it essentially acquires data that are difficult for the language modeling task, thus favoring data with a more diverse vocabulary. All other methods acquire similarly diverse data, except BADGE that has the lowest score. Interestingly, we observe a different pattern when evaluating diversity in the model feature space (using the `[CLS]` representations). BERTKM has the highest DIV.-F score, as expected, while CAL and ENTROPY have the lowest. This supports our hypothesis that uncertainty sampling tends to acquire uncertain but similar examples, while CAL by definition constrains its search in similar examples in the feature space that lie close to the decision boundary (contrastive examples). As for uncertainty, we observe that ENTROPY and CAL acquire the most uncertain examples, with average entropy almost twice as high as all other methods. Finally, regarding representativeness of the acquired batches, we see that CAL obtains the highest score, followed by ENTROPY, with the rest AL strategies to acquire less representative data.

Overall, our analysis validates assumptions on the properties of data expected to be selected by the various acquisition functions. Our findings show that diversity in the raw text does not necessarily correlate with diversity in the feature space. In other words, low DIV.-F does not translate to low diversity in the distribution of acquired tokens (DIV.-I), suggesting that CAL can acquire similar examples in the feature space that have sufficiently diverse inputs. Furthermore, combining the results of our AL experiments (Figure 2) and our analysis (Table 3) we conclude that the best performance of CAL, followed by ENTROPY, is due to acquiring uncertain data. We observe that the most notable difference, in terms of selected data, between the two approaches and the rest is uncertainty (UNC.), suggesting perhaps the superiority of uncertainty over diversity sampling. We show that CAL improves over ENTROPY because our algorithm "guides" the focus of uncertainty sampling by not considering redundant uncertain data that lie away from the decision boundary and thus improving representativeness. We finally find that RANDOM is evidently the worst approach, as it selects the least diverse and uncertain data on average compared to all methods.

7 Related Work

Uncertainty Sampling Uncertainty-based acquisition for AL focuses on selecting data points that the model predicts with low confidence. A simple uncertainty-based acquisition function is *least confidence* (Lewis and Gale, 1994) that sorts data in descending order from the pool by the probability of not predicting the most confident class. Another approach is to select samples that maximize the predictive entropy. Houlsby et al. (2011)

propose Bayesian Active Learning by Disagreement (BALD), a method that chooses data points that maximize the mutual information between predictions and model's posterior probabilities. Gal et al. (2017) applied BALD for deep neural models using Monte Carlo dropout (Gal and Ghahramani, 2016) to acquire multiple uncertainty estimates for each candidate example. Least confidence, entropy and BALD acquisition functions have been applied in a variety of text classification and sequence labeling tasks, showing to substantially improve data efficiency (Shen et al., 2017; Siddhant and Lipton, 2018; Lowell and Lipton, 2019; Kirsch et al., 2019; Shelmanov et al., 2021; Margatina et al., 2021).

Diversity Sampling On the other hand, diversity or representative sampling is based on selecting batches of unlabeled examples that are representative of the unlabeled pool, based on the intuition that a representative set of examples once labeled, can act as a surrogate for the full data available. In the context of deep learning, Geifman and El-Yaniv (2017) and Sener and Savarese (2018) select representative examples based on core-set construction, a fundamental problem in computational geometry. Inspired by generative adversarial learning, Gissin and Shalev-Shwartz (2019) define AL as a binary classification task with an adversarial classifier trained to not be able to discriminate data from the training set and the pool. Other approaches based on adversarial active learning, use out-of-the-box models to perform adversarial attacks on the training data, in order to approximate the distance from the decision boundary of the model (Ducoffe and Precioso, 2018; Ru et al., 2020).

Hybrid There are several existing approaches that combine representative and uncertainty sampling. Such approaches include active learning algorithms that use meta-learning (Baram et al., 2004; Hsu and Lin, 2015) and reinforcement learning (Fang et al., 2017; Liu et al., 2018), aiming to learn a policy for switching between a diversity-based or an uncertainty-based criterion at each iteration. Recently, Ash et al. (2020) propose Batch Active learning by Diverse Gradient Embeddings (BADGE) and Yuan et al. (2020) propose Active Learning by Processing Surprisal (ALPS), a cold-start acquisition function specific for pretrained language models. Both methods construct representations for the unlabeled data based on uncertainty, and then use them for clustering; hence combining both uncertainty and diversity sampling. The effectiveness of AL in a variety of NLP tasks with pretrained language models, e.g. BERT (Devlin et al., 2019), has empirically been recently evaluated by Ein-Dor et al. (2020), showing substantial improvements over random sampling.

8 Conclusion & Future Work

We present CAL, a novel acquisition function for AL that acquires *contrastive examples*; data points which are similar in the model feature space and yet the model outputs maximally different class probabilities. Our approach uses information from the feature space to create neighborhoods for each unlabeled example, and predictive likelihood for ranking the candidate examples. Empirical experiments on various in-domain and out-of-domain scenarios demonstrate that CAL performs better than other acquisition functions in the majority of cases. After analyzing the actively acquired datasets obtained with all methods considered, we conclude that entropy is the hardest baseline to beat, but our approach improves it by guiding uncertainty sampling in regions near the decision boundary with more informative data.

Still, our empirical results and analysis show that there is no single acquisition function to outperform all others consistently *by a large margin*. This demonstrates that there is still room for improvement in the AL field.

Furthermore, recent findings show that in specific tasks, as in Visual Question Answering (VQA), complex acquisition functions might not outperform random sampling because they tend to select *collective outliers* that hurt model performance (Karamcheti et al., 2021). We believe that taking a step back and analyzing the behavior of standard acquisition functions, e.g. with Dataset Maps (Swayamdipta et al., 2020), might be beneficial. Especially, if similar behavior appears in other NLP tasks too.

Another interesting future direction for CAL, related to interpretability, would be to evaluate whether acquiring contrastive examples for the *task* (Kaushik et al., 2020; Gardner et al., 2020) is more beneficial than contrastive examples for the *model*, as we do in CAL.

Acknowledgments

KM and NA are supported by Amazon through the Alexa Fellowship scheme.

References

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2020. Deep batch active learning by diverse, uncertain gradient lower bounds. In *Proceedings of the International Conference on Learning Representations*.

Yoram Baram, Ran El-Yaniv, and Kobi Luz. 2004. Online choice of active learning algorithms. *Journal of Machine Learning Research*, 5:255–291.

Zalán Bodó, Zsolt Minier, and Lehel Csató. 2011. Active learning with clustering. In *Proceedings of the Active Learning and Experimental Design workshop In conjunction with AISTATS 2010*, volume 16, pages 127–139.

Klaus Brinker. 2003. Incorporating diversity in active learning with support vector machines. In *Proceedings of the International Conference on Machine Learning*, pages 59–66.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning*, volume 119, pages 1597–1607.

David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. 1996. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4(1):129–145.

Sanjoy Dasgupta. 2011. Two faces of active learning. *Theoretical Computer Science*, 412(19):1767–1781. Algorithmic Learning Theory (ALT 2009).

Franck Dernoncourt and Ji Young Lee. 2017. PubMed 200k RCT: a dataset for sequential sentence classification in medical abstracts. In *Proceedings of the Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 308–313.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 295–302.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah A. Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *ArXiv*.

Melanie Ducoffe and Frederic Precioso. 2018. Adversarial active learning for deep networks: a margin based approach.

Liat Ein-Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2020. Active learning for BERT: An empirical study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 7949–7962.

Meng Fang, Yuan Li, and Trevor Cohn. 2017. Learning how to active learn: A deep reinforcement learning approach. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 595–605.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the International Conference on Machine Learning*, volume 48, pages 1050–1059.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep Bayesian active learning with image data. In *Proceedings of the International Conference on Machine Learning*, volume 70, pages 1183–1192.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323.

Yonatan Geifman and Ran El-Yaniv. 2017. Deep active learning over the long tail. *CoRR*, abs/1711.00941.

Daniel Gissin and Shai Shalev-Shwartz. 2019. Discriminative active learning. *CoRR*, abs/1907.06347.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2021. Supervised contrastive learning for pre-trained language model fine-tuning. In *Proceedings of the International Conference on Learning Representations*.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020. Pretrained transformers improve out-of-distribution robustness. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 2744–2751.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian active learning for classification and preference learning. *ArXiv*.

Wei-Ning Hsu and Hsuan-Tien Lin. 2015. Active learning by learning. In *Proceedings of the Conference of the Association for the Advancement of Artificial Intelligence*, pages 2659–2665.

Siddharth Karamcheti, Ranjay Krishna, Li Fei-Fei, and Christopher Manning. 2021. Mind your outliers! investigating the negative impact of outliers on active learning for visual question answering. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, pages 7265–7281.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *Proceedings of the International Conference on Learning Representations*.

Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. 2019. BatchBALD: Efficient and diverse batch acquisition for deep bayesian active learning. In *Proceedings of the Conference on Neural Information Processing Systems*, pages 7026–7037.

Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. 2017. A continuously growing dataset of sentential paraphrases. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1224–1234.

David D. Lewis and Jason Catlett. 1994. Heterogeneous uncertainty sampling for supervised learning. In *Machine Learning Proceedings 1994*, pages 148–156.

David D. Lewis and William A. Gale. 1994. A sequential algorithm for training text classifiers. In *In Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Ming Liu, Wray Buntine, and Gholamreza Haffari. 2018. Learning how to actively learn: A deep imitation learning approach. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1874–1883.

David Lowell and Zachary C Lipton. 2019. Practical obstacles to deploying active learning. *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, pages 21–30.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150.

Katerina Margatina, Loïc Barrault, and Nikolaos Aletras. 2021. Bayesian active learning with pretrained language models. *CoRR*, abs/2104.08320.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the International Conference on Neural Information Processing Systems*, page 3111–3119.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035.

Nicholas Roy and Andrew McCallum. 2001. Toward optimal active learning through sampling estimation of error reduction. In *Proceedings of the International Conference on Machine Learning*, page 441–448.

Dongyu Ru, Jiangtao Feng, Lin Qiu, Hao Zhou, Mingxuan Wang, Weinan Zhang, Yong Yu, and Lei Li. 2020. Active sentence learning by adversarial uncertainty sampling in discrete space. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4908–4917, Online. Association for Computational Linguistics.

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. In *Proceedings of the International Conference on Learning Representations*.

Burr Settles. 2009. Active learning literature survey. Computer sciences technical report.

Artem Shelmanov, Dmitri Puzyrev, Lyubov Kupriyanova, Denis Belyakov, Daniil Larionov, Nikita Khromov, Olga Kozlova, Ekaterina Artemova, Dmitry V. Dylov, and Alexander Panchenko. 2021. Active learning for sequence tagging with deep pre-trained models and Bayesian uncertainty estimates. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, pages 1698–1712.

Dan Shen, Jie Zhang, Jian Su, Guodong Zhou, and Chew-Lim Tan. 2004. Multi-criteria-based active learning for named entity recognition. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 589–596.

Yanyao Shen, Hyokun Yun, Zachary Lipton, Yakov Kronrod, and Animashree Anandkumar. 2017. Deep active learning for named entity recognition. In *Proceedings of the Workshop on Representation Learning for NLP*, pages 252–256.

Aditya Siddhant and Zachary C Lipton. 2018. Deep bayesian active learning for natural language processing: Results of a Large-Scale empirical study.

In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 2904–2909.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642.

Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 9275–9293.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Michelle Yuan, Hsuan-Tien Lin, and Jordan Boyd-Graber. 2020. Cold-start active learning through self-supervised language modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7935–7948, Online. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28, pages 649–657. Curran Associates, Inc.

Fedor Zhdanov. 2019. Diverse mini-batch active learning.

Jingbo Zhu, Huizhen Wang, Tianshun Yao, and Benjamin K Tsou. 2008. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In *Proceedings of the International Conference on Computational Linguistics*, pages 1137–1144.

A Appendix

A.1 Data & Hyperparameters

In this section we provide details of all the datasets we used in this work and the hyperparparameters used for training the model. For QNLI, IMDB and SST-2 we randomly sample 10% from the training set to serve as the validation set, while for AG-NEWS and QQP we sample 5%. For the DBPEDIA dataset we undersample both training and validation datasets (from the standard splits) to facilitate our AL simulation (i.e. the original dataset consists of 560K training and 28K validation data examples). For all datasets we use the standard test set, apart from SST-2, QNLI and QQP datasets that are taken from the GLUE benchmark (Wang et al., 2019) we use the development set as the held-out test set and subsample a development set from the training set.

For all datasets we train BERT-BASE (Devlin et al., 2019) from the HuggingFace library (Wolf et al., 2020) in Pytorch (Paszke et al., 2019). We train all models with batch size 16, learning rate $2e-5$, no weight decay, AdamW optimizer with epsilon $1e-8$. For all datasets we use maximum sequence length of 128, except for IMDB that contain longer input texts, where we use 256. To ensure reproducibility and fair comparison between the various methods under evaluation, we run all experiments with the same five seeds that we randomly selected from the range $[1, 9999]$. We evaluate the model 5 times per epoch on the development set following Dodge et al. (2020) and keep the one with the lowest validation loss. We use the code provided by Yuan et al. (2020) for ALPS, BADGE and BERTKM.

A.2 Efficiency

In this section we compare the efficiency of the acquisition functions considered in our experiments. We denote m the number of labeled data in $\mathcal{D}_{\textbf{lab}}$, n the number of unlabeled data in $\mathcal{D}_{\textbf{pool}}$, C the number of classes in the downstream classification task, d the dimension of embeddings, t is fixed number of iterations for k-MEANS, l the maximum sequence length and k the acquisition size. In our experiments, following (Yuan et al., 2020), $k = 100$, $d = 768$, $t = 10$, and $l = 128$[12]. ALPS requires $\mathcal{O}(tknl)$ considering that the surprisal embeddings are computed. BERTKM and BADGE, the most computationally heavy approaches, require $\mathcal{O}(knd)$ and $\mathcal{O}(Cknd)$ respectively, given that gradient embeddings are computed for BADGE [13]. On the other hand, ENTROPY only requires n forward passes though the model, in order to obtain the logits for all the data in $\mathcal{D}_{\textbf{pool}}$. Instead, our approach, CAL, first requires $m + n$ forward passes, in order to acquire the logits and the CLS representations of the the data (in $\mathcal{D}_{\textbf{pool}}$ and $\mathcal{D}_{\textbf{lab}}$) and then one iteration for all data in $\mathcal{D}_{\textbf{pool}}$ to obtain the scores.

We present the runtimes in detail for all datasets and acquisition functions in Tables 4 and 5. First, we define the total *acquisition time* as a sum of two types of times; *inference* and *selection* time. Inference time is the time that is required in order to pass all data from the model to acquire predictions or probability distributions or model encodings (representations). This is explicitly required for the uncertainty-based methods, like ENTROPY, and our method CAL. The remaining time is considered *selection* and essentially is the time for all necessary computations in order to rank and select the b most important examples from $\mathcal{D}_{\textbf{pool}}$.

We observe in Table 4 that the diversity-based functions do not require this explicit inference time, while for ENTROPY it is the only computation that is needed (taking the argmax of a list of uncertainty scores is negligible). CAL requires both inference and selection time. We can see that inference time of CAL is a bit higher than ENTROPY because we do $m+n$ forward passes instead of n, that is equivalent to both $\mathcal{D}_{\textbf{pool}}$ and $\mathcal{D}_{\textbf{lab}}$ instead of only $\mathcal{D}_{\textbf{pool}}$. The selection time for CAL is the *for-loop* as presented in our Algorithm 1. We observe that it is often less computationally expensive than the inference step (which is a simple forward pass through the model). Still, there is room for improvement in order to reduce the time complexity of this step.

In Table 5 we present the total time for all datasets (ordered with increasing $\mathcal{D}_{\textbf{pool}}$ size) and the average time for each acquisition function, as a means to rank their efficiency. Because we do not apply all acquisition functions to all datasets we compute three different average scores in order to ensure fair comparison. AVG.-ALL is the average time across all 7 datasets and is used to compare RANDOM, ALPS, ENTROPY and CAL. AVG.-3 is the average time across the first 3 datasets (IMDB, SST-2 and DBPEDIA) and is used to compare all

[12] Except for IMDB where $l = 256$.

[13] This information is taken from Section 6 of Yuan et al. (2020).

	DBPEDIA	IMDB	SST-2	QNLI	AGNEWS	PUBMED	QQP
RANDOM	$(0, 0)$	$(0, 0)$	$(0, 0)$	$(0, 0)$	$(0, 0)$	$(0, 0)$	$(0, 0)$
ALPS	$(0, 181)$	$(0, 222)$	$(0, 733)$	$(0, 1607)$	$(0, 2309)$	$(0, 5878)$	$(0, 14722)$
BERTKM	$(0, 467)$	$(0, 431)$	$(0, 4265)$	$(0, 8138)$	$(0, 9344)$	$(0, 25965)$	$(-, -)$
BADGE	$(0, 12871)$	$(0, 3816)$	$(0, 25640)$	$(-, -)$	$(-, -)$	$(-, -)$	$(-, -)$
ENTROPY	$(103, 1)$	$(107, 0)$	$(173, 0)$	$(331, 0)$	$(402, 0)$	$(596, 0)$	$(1070, 0)$
CAL	$(133, 49)$	$(212, 61)$	$(464, 244)$	$(528, 376)$	$(656, 628)$	$(1184, 1445)$	$(1541, 2857)$

Table 4: Runtimes (in seconds) for all datasets and acquisition functions. In each cell of the table we present a tuple (i, s) where i is the *inference time* and s the *selection time*. *Inference time* is the time for the model to perform a forward pass for all the unlabeled data in $\mathcal{D}_{\textbf{pool}}$ and *selection time* is the time that each acquisition function requires to rank all candidate data points and select b for annotation (for a single iteration). Since we cannot report the runtimes for *every* model in the AL pipeline (at each iteration the size of $\mathcal{D}_{\textbf{pool}}$ changes), we provide the median.

	DBPEDIA	IMDB	SST-2	QNLI	AGNEWS	PUBMED	QQP	AVG.-ALL	AVG.-3	AVG.-6
RANDOM	0	0	0	0	0	0	0	0	0	0
ALPS	181	222	733	1607	2309	5878	14722	3664	378	1821
BERTKM	467	431	4265	8138	9344	25965	–	–	1721	8101
BADGE	12871	3816	25640	–	–	–	–	–	14109	–
ENTROPY	104	107	173	331	402	596	1070	397	128	285
CAL	182	273	708	904	1284	2629	4398	1482	387	996

Table 5: Runtimes (in seconds) for all datasets and acquisition functions. In each cell of the table we present the total acquisition time (inference and selection). AVG.-ALL shows the average acquisition time for each acquisition function for all datasets, AVG.-6. for all datasets except QQP and AVG.-3 for the 3 first datasets only (DBPEDIA, IMDB, SST-2).

acquisition functions. Finally, AVG.-6 is the average time across all datasets apart from QQP and is used to compare RANDOM, ALPS, BERTKM, ENTROPY and CAL.

We first observe that ENTROPY is overall the most efficient acquisition function. According to the AVG.-ALL column, we observe that CAL is the second most efficient function, followed by ALPS. According to the AVG.-6 we observe the same pattern, with BERTKM to be the slowest method. Finally, we compare all acquisition functions in the 3 smallest (in terms of size of $\mathcal{D}_{\textbf{pool}}$) datasets and find that ENTROPY is the fastest method followed by ALPS and CAL that require almost 3 times more computation time. The other clustering methods, BERTKM and BADGE, are significantly more computationally expensive, requiring respectively 13 and 100(!) times more time than ENTROPY.

Interestingly, we observe the effect of the acquisition size (2% of $\mathcal{D}_{\textbf{pool}}$ in our case) and the size of $\mathcal{D}_{\textbf{pool}}$ in the clustering methods. As these parameters increase, the computation of the corresponding acquisition function increases dramatically. For example, we observe that in the 3 smallest datasets that ALPS requires similar time to CAL. However, when we increase b and m (i.e. as we move from DBPEDIA with $20K$ examples in $\mathcal{D}_{\textbf{pool}}$ to QNLI with $100K$ etc - see Table 1) we observe that the acquisition time of ALPS becomes twice as much as that of CAL. For instance, in QQP with acquisition size 3270 we see that ALPS requires 14722 seconds on average, while CAL 4398. This shows that even though our approach is more computationally expensive as the size of $\mathcal{D}_{\textbf{pool}}$ increases, the complexity is linear, while for the other hybrid methods that use clustering, the complexity grows exponentially.

A.3 Reproducibility

All code for data preprocessing, model implementations, and active learning algorithms is made available at https://github.com/mourga/contrastive-active-learning. For questions regarding the implementation, please contact the first author.

3.3 Impact

According to Google Scholar, the paper has received 140 citations as of May 2024 and it was featured in numerous surveys (e.g. Tsvigun et al., 2022; Zhang et al., 2022f; Treviso et al., 2023; Schröder et al., 2023; Rauch et al., 2023; Hu et al., 2023b,a; Zhang, 2023; Ghose and Nguyen, 2024).

CAL has directly influenced several followup studies (e.g. Zhang et al., 2022c; Maekawa et al., 2022; Yu et al., 2022; Azeemi et al., 2023b; Hassan and Alikhani, 2023). Zhang et al. (2022c) propose ALLSH: Active Learning Guided by Local Sensitivity and Hardness. Similar to CAL, the motivation is to retrieve unlabeled samples with a local sensitivity and hardness-aware acquisition function. ALLSH generates data copies through local perturbations and selects data points whose predictive likelihoods diverge the most from their copies.

The analysis we performed at our paper has also been used by several studies related to data selection methods. Specifically, the analysis presented in Section 5 of our paper, that uses diversity (DIV-I, DIV-F) and representativeness (REPR.) metrics, has been established as standard practice for comprehensive analysis in the following, but not limited to, papers (Su et al., 2023; Wan et al., 2024).

Several recent works perform extensive experimental studies comparing popular acquisition functions for active learning, for various NLP tasks, including CAL as a benchmark acquisition function (e.g. Zhang et al., 2022b,a; Romberg and Escher, 2022; Karisani et al., 2022; Steegh and Sileno, 2023; Zhang et al., 2022c; Maekawa et al., 2022; Yu et al., 2022, 2023; Li and Qiu, 2023; Köksal et al., 2023; Zeng and Zubiaga, 2023; Deng et al., 2023; Chen et al., 2023; Le et al., 2024; Chen et al., 2024; Pecher et al., 2024; Ying et al., 2024). Zhang et al. (2022c,b); Romberg and Escher (2022); Maekawa et al. (2022); Le et al. (2024), among several others, show that CAL outperforms the standard maximum entropy baseline, while Pecher et al. (2024) even shows that CAL has considerably lower standard deviation across multiple runs.

CAL has been applied successfully even beyond the NLP domain (Wan et al., 2024; Ying et al., 2024). Garg and Roy (2023) showed that CAL was among the top performing baselines in the CIFAR-10 and CIFAR-100 datasets in computer vision, while Azeemi et al. (2023a) used CAL as an inspiration for their COWERAGE algorithm for representative subset selection in self-supervised Automatic Speech Recognition (ASR).

3.4 Discussion

Despite its simplicity, CAL outperformed other uncertainty-based and diversity based algorithms in our experiments. After analyzing the actively acquired datasets obtained with all methods considered, we concluded that entropy was the hardest baseline to beat, but our approach improved it by guiding uncertainty sampling in regions near the decision boundary with more informative data.

One very interesting discovery is shown in the Ablation section of the paper (Figure 3) where we compare many versions of CAL (with different settings). We included as a baseline the "opposite CAL", where instead of choosing the data points that scored the highest KL divergence with their labeled neighbours we chose the ones with the lowest score (orange line in the figure). We found that not only this version did not work well, as expected, but even degraded the AL performance. The model performed better when trained with randomly chosen 2% of the available data instead of up to 15% of the data chosen with this approach. This very clearly shows that *not all data is equal*, that including only easy, redundant data points can dramatically harm performance, and further solidifying that hard uncertain examples are imperative for a high-performing data-efficient model, which is the main takeaway of the paper.

One limitation of CAL is that we do not take into account a specific filter to avoid selecting redundant data points. We do guide selection with our neighbourhood formulation, but still we do not further investigate how similar are the data points that are to be selected in the batch. Adding another criterion there would most likely further improved CAL's performance and sharpened its superiority to entropy sampling. Another important limitation, not of CAL but in general of AL is the fragility of the ranking of the acquisition strategies based on the setting they are applied. There is no universally superior algorithm that outperforms all others. This is evident by the results, where the performance lines in the plots are close to each other in many cases. We opted to include the standard deviations as well in order to provide a clear presentation of the results and avoid any misleading interpretation. This limitation is also highlighted by the literature (Margatina and Aletras, 2023; Ghose and Nguyen, 2024).

Chapter 4

Publication III: Active In-Context Learning Learning with Large Language Models

The main contribution of this chapter is the paper *Active Learning Principles for In-Context Learning with Large Language Models*, which was published at the *Findings of the Association for Computational Linguistics* at the Empirical Methods of Natural Language Processing conference in December 2023. We first outline the motivation for this work (Section 4.1), followed by the paper itself (Section 4.2), the impact that it has had so far (Section 4.3), and discussion (Section 4.4).

4.1 Introduction

Amidst the dynamic advancements in the NLP field, a notable shift from the conventional supervised learning paradigm to few-shot learning, also known as in-context learning, has emerged. Rather than fine-tuning smaller language models like BERT (Devlin et al., 2019), the focus has turned towards leveraging larger and more capable Language Model Models (LLMs) such as GPT (Radford et al., 2019; Brown et al., 2020; Black et al., 2022; OpenAI, 2023) and OPT (Zhang et al., 2022d) models, boasting millions or billions of parameters. These models can be utilized off-the-shelf, without requiring any weight updates, through prompting with in-context examples (i.e., demonstrations) serving as the new learning paradigm. Our interest lies in exploring how active learning (AL) algorithms, previously analyzed within the supervised learning framework, could adapt to these models. Thus, our research shifted towards investigating whether AL acquisition functions could serve as effective data selection

algorithms for in-context learning.

In this study, our primary goal was to redefine the notion of data efficiency within the context of in-context learning, drawing inspiration from conventional AL methodologies. We approached this by aiming to identify a subset of k examples from a pool of labeled or unlabeled data, which would act as demonstrations to an LLM, ultimately maximizing performance on a held-out test set. To achieve this, we explored the effectiveness of prevalent AL approaches centered around uncertainty (Lewis and Gale, 1994; Cohn et al., 1996; Gal et al., 2017), diversity (Brinker, 2003; Bodó et al., 2011; Sener and Savarese, 2018) and similarity (Margatina et al., 2021; Kirsch et al., 2021; Liu et al., 2022), as methods for selecting demonstrations in the context of in-context learning. Our evaluation spanned fifteen models ranging from 125M to 30B parameters within the GPT and OPT families, across fifteen classification and nine multi-choice tasks. Our findings underscored the superiority of selecting in-context examples that closely aligned with the semantic content of the input test examples across all tasks, model families, and sizes, while also highlighting nuances in the efficacy of uncertainty sampling between supervised and in-context learning settings. Moreover, we observed that larger models may exhibit greater benefits from uncertain demonstrations, indicating a potential interplay between model size and the utilization of uncertainty as an emerging capability in LLMs.

4.2 The Paper

Author Contributions

The paper is co-authored by myself, Timo Schick, Nikolaos Aletras and Jane Yu. As the lead author, I conceived the idea to explore this research question, implemented the codebase, performed the experiments, and wrote the paper. This paper was fulfilled during my internship at Meta AI (FAIR), where Timo Schick and Jane Yu were my advisors. They participated in many discussions and proofread the paper. Nikolaos Aletras helped with brainstorming the original idea, offered suggestions and helped revise the final version.

Active Learning Principles for In-Context Learning with Large Language Models

Katerina Margatina[◇*] **Timo Schick**[†] **Nikolaos Aletras**[◇] **Jane Dwivedi-Yu**[†]
[◇]University of Sheffield [†]FAIR, Meta
{k.margatina, n.aletras}@sheffield.ac.uk
janeyu@meta.com

Abstract

The remarkable advancements in large language models (LLMs) have significantly enhanced predictive performance in few-shot learning settings. By using only a small number of labeled examples, referred to as demonstrations, LLMs can effectively perform the task at hand through in-context learning. However, the process of selecting demonstrations for maximizing performance has received limited attention in prior work. This paper addresses the issue of identifying the most informative demonstrations for few-shot learning by approaching it as a pool-based Active Learning (AL) problem over a single iteration. We compare standard AL algorithms based on uncertainty, diversity, and similarity, and consistently observe that the latter outperforms all other methods, including random sampling. Our extensive experimentation involving a diverse range of GPT and OPT models across 24 classification and multi-choice tasks, coupled with thorough analysis, unambiguously demonstrates the importance of using demonstrations that are semantically similar to the domain of the test examples. In fact, we show higher average classification performance using "similar" demonstrations with GPT-2 (124M) than random demonstrations with GPT-Neox (20B). Notably, while diversity sampling shows promise, uncertainty sampling, despite its success in conventional supervised learning AL scenarios, performs poorly in in-context learning.

1 Introduction

The field of Natural Language Processing (NLP) has recently witnessed a remarkable paradigm shift with the emergence of in-context learning with large language models (LLMs), also referred to as few-shot learning (Brown et al., 2020). Traditionally, NLP systems heavily relied on supervised learning approaches, where large amounts of labeled training data were necessary to achieve high predictive performance. However, in-context learning has changed this status-quo by enabling LLMs to learn from limited, context-specific examples and adapt to new tasks and domains with remarkable proficiency (Zhao et al., 2021; Chowdhery et al., 2022; García et al., 2023; Wei et al., 2023b; Touvron et al., 2023; Bubeck et al., 2023). Unlike more traditional approaches, which require extensive retraining or fine-tuning for every new task, in-context learning empowers LLMs to generalize from a few examples that are fed to the model through prompting to learn a new task at hand, without any weight updates.

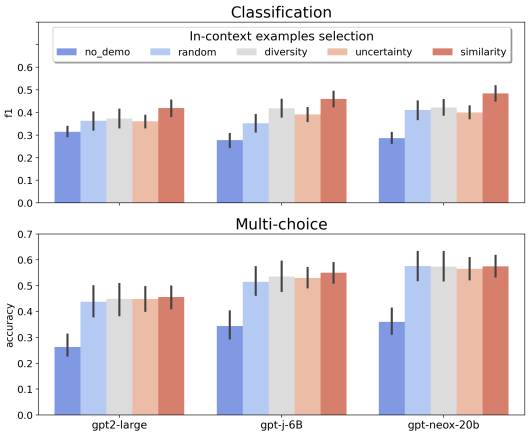


Figure 1: Performance of different in-context selection algorithms in classification and multi-choice tasks.

The data efficiency of few-shot in-context learning of LLMs is indeed remarkable with only a small number of demonstrations.[1] Still, such demonstrations constitute *labeled* data examples, raising two key questions: (1) When faced with tasks where there is only *unlabeled* data available, how can we select the most appropriate samples to label and then use as in-context demonstrations? (2) When we have *labeled* data for a given task, how can

[*] Work done during an internship at FAIR, Meta.

[1] We use the terms *in-context examples*, *few-shot examples*, *demonstrations*, *descriptors* and *exemplars* interchangeably throughout the paper.

Findings of the Association for Computational Linguistics: EMNLP 2023, pages 5011–5034
December 6-10, 2023

we efficiently identify the most informative combination of demonstrations for in-context learning? Answering these questions is essential to ensure effective and efficient few-shot learning using LLMs.

A growing line of work has investigated how in-context learning works (Reynolds and McDonell, 2021; Razeghi et al., 2022; Xie et al., 2022; Ye et al., 2023b), which demonstrations to use (Liu et al., 2022; Zhang et al., 2022b; Wu et al., 2022; Kim et al., 2022), how to form the prompt (Zhao et al., 2021; Lu et al., 2022; Yang et al., 2023) and whether ground truth labels matter (Webson and Pavlick, 2022; Min et al., 2022; Yoo et al., 2022; Wang et al., 2022; Wei et al., 2023b). Still, to the best of our knowledge, no prior work has explored the problem of in-context demonstration selection explicitly through the lens of active learning (AL).

Based on the core principle that not all data points are equally useful, AL (Cohn et al., 1996; Settles, 2009) aims to identify the most informative instances from a pool of unlabeled data for annotation. Iterating through model training, data acquisition and human annotation, the goal is to achieve data efficiency. A data-efficient AL algorithm ensures that a model achieves satisfactory performance on a withheld test set by selecting only a small fraction of the unlabeled data for annotation that typically is better than randomly selecting and annotating data of equal size.

In this paper, our main aim is to redefine the concept of data efficiency within the framework of in-context learning inspired by conventional active learning settings. For this purpose, we assume that given a pool of labeled or unlabeled data, the objective is to identify a set of k examples that will serve as demonstrations to an LLM, resulting in optimal performance on a held-out test set. Given this formulation of data efficiency, we explore the effectiveness of the most prevalent AL approaches based on uncertainty (Lewis and Gale, 1994; Cohn et al., 1996; Gal et al., 2017), diversity (Brinker, 2003; Bodó et al., 2011; Sener and Savarese, 2018) and similarity (Margatina et al., 2021; Kirsch et al., 2021; Liu et al., 2022), as demonstration selection methods for in-context learning (Figure 1).

Our key contributions are as follows:

- We formulate the selection of in-context examples as a single iteration AL problem and explore the effectiveness of four standard approaches: *uncertainty*, *diversity*, *similarity* and *random* sampling.
- We evaluate 15 models, between 125M and 30B parameters, from the GPT (Radford et al., 2019; Brown et al., 2020; Black et al., 2022) and OPT (Zhang et al., 2022a) families in 15 classification and 9 multi-choice tasks, using different AL sampling techniques to select demonstrations for few-shot learning.
- We demonstrate that while diversity and uncertainty sampling perform slightly better than random sampling, choosing in-context examples that are semantically similar to the input test examples outperforms consistently all other methods by a large margin across model families and sizes in all tasks.
- We show that while uncertainty sampling is one of the strongest AL approaches in supervised learning, this does not generalize to in-context learning, where interestingly it underperforms. Our analysis, however, shows that larger models might perform better with uncertain demonstrations, hinting that uncertainty might be an emerging LLM ability.

2 Active In-context Learning

2.1 Problem Formulation

To build our in-context learning framework with actively acquired demonstrations, depicted in Figure 2, we borrow the formulation from the standard pool-based active learning paradigm. We consider an AL setting where we have a large pool of unlabeled data from which we want to sample a batch of k data points using a data acquisition algorithm. We assume that these k are subsequently labeled by humans (Figure 2, top). Instead of following the standard approach that involves multiple iterations of data selection and model training, we only perform a single iteration (Longpre et al., 2022), since we do not train or perform any model-in-the-loop updates. We use the acquired set of k examples as demonstrations for in-context learning with an LLM (i.e., as part of the prompt). We assume the existing datasets as the pool from which to select these k examples. The goal is to find the most informative examples from the pool, which are expected to yield improved performance on the test set when employed as a few-shot prompt, compared to demonstrations randomly sampled from the same pool. The resulting prompt consists of the concatenation of the k acquired examples (text

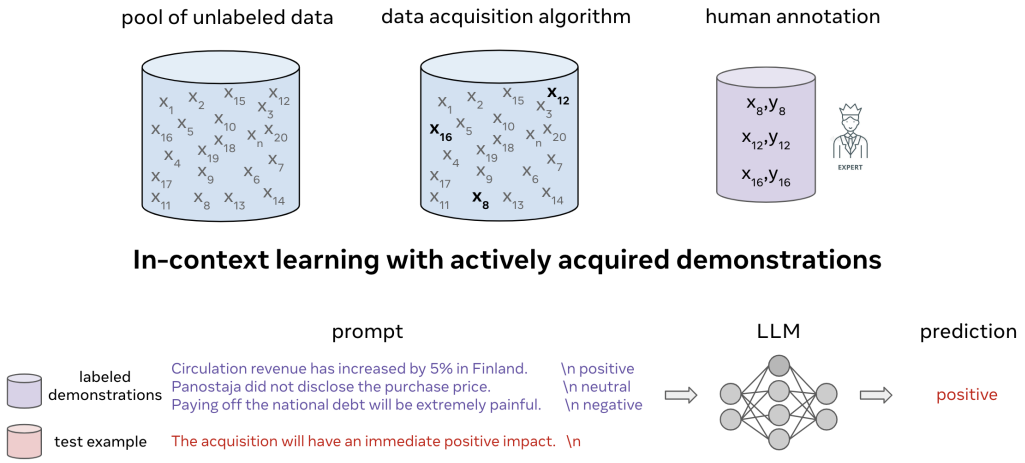


Figure 2: Top: Active data collection (single iteration). Bottom: Prompt construction and model inference.

inputs and labels with standard verbalizers), alongside the test example, repeated for all data instances in the test set (Figure 2, bottom).

2.2 Few-shot Data Acquisition Algorithms

We build few-shot data acquisition algorithms inspired by the most prevalent AL algorithmic families that are `uncertainty` sampling, `diversity` sampling and `similarity` (also known as test-aware sampling) (Zhang et al., 2022c). We acknowledge that there are more elaborate demonstration selection methods for in-context learning that are not considered in our experiments, such as Q-learning (Zhang et al., 2022b), Self Adaptive (Wu et al., 2022), SG-ICL (Kim et al., 2022), MI (Sorensen et al., 2022), *inter alia*. These methods fall beyond the scope of our analysis, as our objective is to gain insights into AL principles for in-context learning, rather than benchmarking all available demonstration sampling algorithms. Additionally, there are techniques, complementary to the aforementioned few-shot data selection methods, such as calibration (Zhao et al., 2021) and prompt re-ordering (Lu et al., 2022), which can further enhance few-shot learning performance, while also being out of the scope of our work.

`Random` The overarching objective of any data selection method, like AL algorithms, is to identify data points that, however used, yield superior models compared to randomly sampled data from the same pool which we consider as a baseline method.

`Diversity` The first data selection method that we use as a representative for the diversity family of methods is a simple clustering technique, similar to Yu et al. (2022). Specifically, we first encode all data points in the pool of unlabeled data with `Sentence-BERT` (Reimers and Gurevych, 2019) embeddings and then we perform k-means clustering.[2] We choose the number of clusters to be k and select one data point from each cluster. The underlying principle of this approach is that leveraging a diverse set of in-context examples can offer greater advantages compared to random sampling. This selection strategy ensures that the chosen demonstrations are likely to encompass a broad range of information, enhancing the overall effectiveness of the learning process.

`Uncertainty` The second approach is an uncertainty-based sampling algorithm that is based on `SPELL`, proposed by Gonen et al. (2022). Since we use an off-the-shelf LLM that does not have a fine-tuned classification layer, we cannot compute the model probabilities associated with each class (for a classification or multi-choice task). This essentially means that we cannot use standard AL uncertainty baselines such as maximum entropy or least confidence. Instead, we can use the loss, i.e., perplexity, of the LLM to score each candidate example from the pool. Gonen et al. (2022) define perplexity of the prompt as the perplexity of the full prompt sequence, including the input itself, and without the label, averaged over $1,000$ examples. Our approach is different since we want to evaluate the perplexity of each in-context example individually. We also do not do the averaging over a thousand examples as we wanted to make the method more general, without

[2]We use the implementation from `https://www.sbert.net/examples/applications/clustering/`.

the need to assume access to that many examples. The underlying principle guiding this approach is the belief that a high perplexity set of in-context examples can yield greater advantages compared to randomly sampling from the dataset (or at least for data efficiency in a supervised learning setting this is proven to enhance the learning process).

`Similarity` Finally, the third AL algorithm we consider is based on `KATE` a kNN-augmented in-context example selection method proposed by Liu et al. (2022). This method retrieves examples from the pool that are semantically-similar to a test query sample. We use `Sentence-BERT` (Reimers and Gurevych, 2019) representations of both the pool and the test set to find the k-nearest neighbours. The rationale behind this approach is that the most similar demonstrations to the test example will best help the model answer the query. We have to highlight, however, that by definition each test example will have a different prompt, as the k most similar demonstrations will be different. This is a crucial limitation of this approach compared to the others, as it assumes that we are able to acquire labels for any in-context example selected from the pool.

3 Experimental Setup

Models We evaluate 15 LLMs in total, 8 models from the GPT (Radford et al., 2019; Brown et al., 2020; Black et al., 2022) and 7 from the OPT (Zhang et al., 2022a) family. We choose our models to span from a few million to tens of billions parameters, as we want to study how the model size affects the effectiveness of in-context example selection methods. All models considered in this work are publicly available.

Tasks & Datasets Following Min et al. (2022), we evaluate all LLMs in 15 classification and 9 multi-choice tasks taken from the `Crossfit` (Ye et al., 2021) benchmark. We provide details for all tasks and datasets considered in the Appendix A.1.

In-context Learning Prompting Unless specified otherwise, we sample k=16 demonstrations, i.e., labeled data, from the pool with each AL method. After collecting the k input-label pairs, we concatenate them all together with the test example that we want to make a prediction for to form the LLM prompt (Figure 2). Our implementation, including prompt verbalizers, is based on those by Min et al. (2022) and Yoo et al. (2022).

4 Results

Figure 3 shows the results on few-shot in-context learning across all data acquisition methods (`random`, `diversity`, `uncertainty` and `similarity`), model families (GPT and OPT) and tasks (classification and multi-choice question answering).[3] Overall, we observe the anticipated trend of performance enhancement with increasing scale, particularly notable in the multi-choice tasks for both OPT and GPT models.

Still, the most remarkable finding is the substantial performance improvement achieved by selecting `similar` in-context examples for few-shot learning, particularly in classification tasks. This observation aligns with the findings reported by Liu et al. (2022), who demonstrated similar patterns in sentiment analysis tasks with GPT-3. Our results indicate that the selection of appropriate demonstrations can hold greater significance than the number of model parameters, at least within the scope of the models evaluated in this study. In multi-choice tasks, `similarity` is also the top-performing acquisition method, while the other three approaches exhibit closely competitive performance.

The data selection method based on `diversity` is consistently the second best approach after `similarity` (with very few exceptions in the multi-choice tasks for OPT models). Even though it is not the top performing method, we can consider that consistently outperforming random sampling is a strong signal that diversity in the demonstrations is a characteristic of effective demonstrations. Levy et al. (2022) explore the setting of compositional generalization, where models are tested on outputs with structures that are absent from the training set and thus selecting similar demonstrations is insufficient. They show that combining diverse demonstrations with in-context learning substantially improves performance for the task of compositional generalization semantic parsing.

Remarkably, `uncertainty` sampling, typically regarded as one of the best approaches for traditional supervised AL (Shen et al., 2017; Margatina et al., 2022; Schröder et al., 2023), exhibits the lowest performance. This finding contradicts the conventional AL principles that suggest selecting a few highly uncertain labeled data points for data efficiency. Similar to our findings, Gonen et al. (2022) explore the performance variabilty of dif-

[3] We provide the results per dataset and model in the Appendix A.2, including the majority vote baseline.

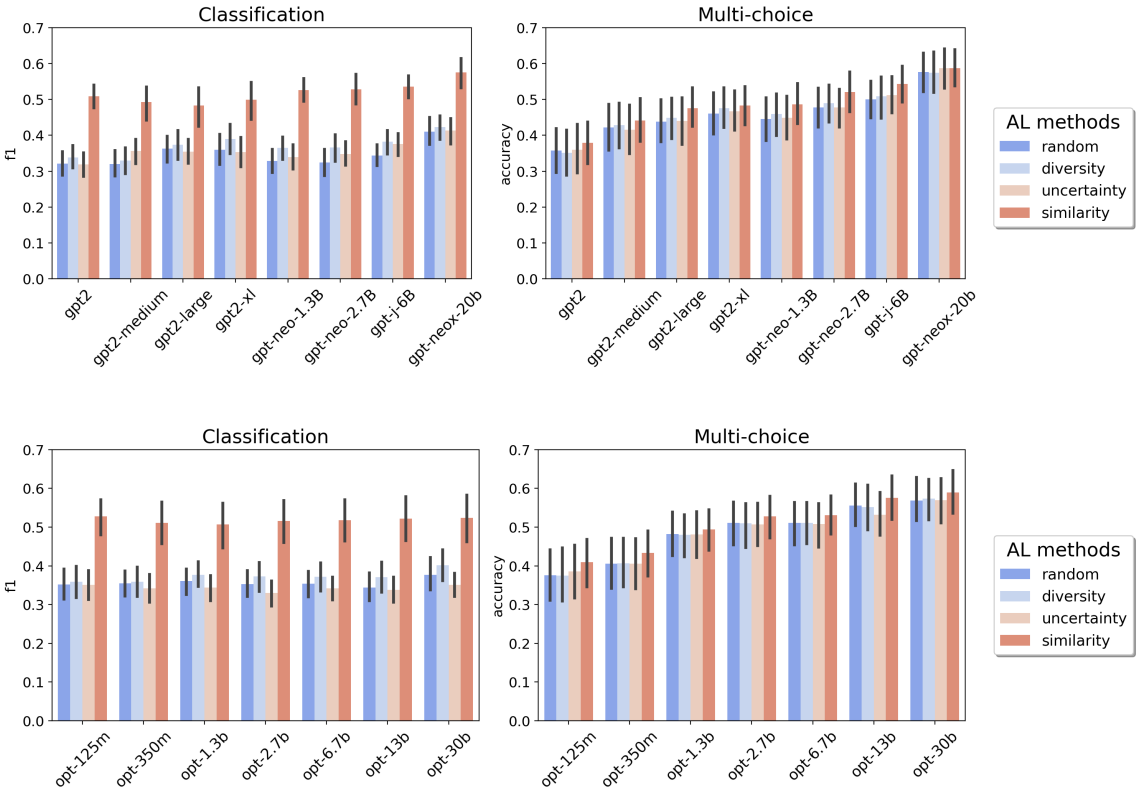


Figure 3: Results for various GPT (top) and OPT (bottom) models and AL methods averaged over 15 classification and 9 multi-choice tasks. *Similarity* is consistently the best performing approach overall, followed by *diversity* and *random*. Interestingly, we observe that *uncertainty* sampling underperforms in this setting of in-context learning.

ferent prompts (consisting of randomly sampled demonstrations) for in-context learning using uncertainty, and find that the lower the perplexity of the prompt is, the better the prompt is able to perform the task. Still, in a later analysis we show that larger models might be able to handle high uncertain prompts better than the smaller ones (§5.4).

5 Analysis

5.1 Effect of Model Size

In order to gain some intuition on the effect of scale, we group together GPT and OPT models with similar number of parameters. We provide the results in Figure 4. Even after aggregating the results from both model families, we do not see any specific pattern as the model parameters increase. We wanted to explore whether the largest models of our collection would behave differently under the varying in-context learning settings, thus perhaps attributing such a behaviour to potential emergent abilities of the bigger LLMs, but we observe the same patterns (in terms of ranking between the considered data selection methods). We believe that this is an interesting avenue of future research, especially as models grow and, most likely, will continue to grow exponentially in terms of model parameters. Our findings show that the in-context learning ability of models from a few millions to a few billions of parameters follows similar patterns. However, this might not be the case when studying even larger models, as primary results hint (Rae et al., 2022; Wei et al., 2023b; Chowdhery et al., 2022; Touvron et al., 2023).

5.2 Ground Truth Demonstrations

We next delve into the debate of whether ground truth demonstrations, i.e., providing the correct label to the in-context examples, is crucial for high performing in-context learning. Various findings have shown mixed results for randomly sampled data, which essentially means that the benefit of ground truth labels depends on the label space or the distribution of inputs specified by the demonstrations (Min et al., 2022; Yoo et al., 2022). In our analysis, we differentiate from prior work by exploring the importance of ground truth demonstrations in the case of leveraging similar in-context examples (§2.2). The rationale is that if the find-

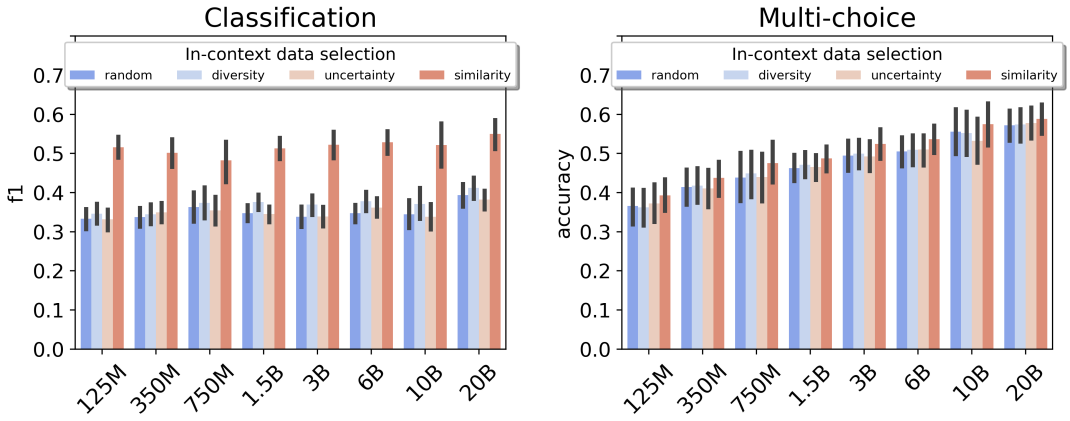


Figure 4: Results per model size.

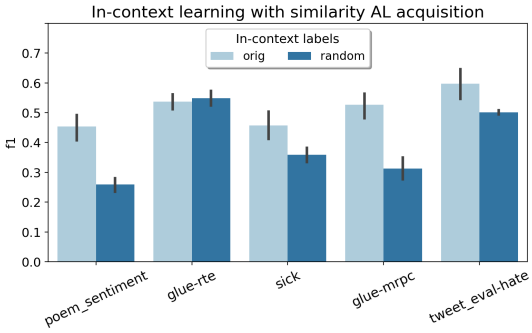


Figure 5: Effect of ground truth labels on in-context learning with with the similarity AL selection method.

ings of Min et al. (2022) ubiquitously hold, then the performance should only marginally drop if we replace ground truth labels with random ones. If the high performance of the `similarity` acquisition method can be retained, we would be able to construct an efficient and effective in-context selection algorithm that would be agnostic to correct labels. However, we find that this is not the case. We show in Figure 5 that for almost all datasets considered in this part of analysis, the performance with random labels drops significantly as expected. There are cases where replacing the original labels with random ones as in Min et al. (2022) retains the same performance (e.g., in the `glue-rte` dataset), but this is certainly a finding that does not generalize overall. In summary, we find that ground truth demonstrations are crucial for high performing, robust in-context learning (Yoo et al., 2022).

5.3 Most vs. Least Similar Demonstrations

To investigate the striking effectiveness of the *similarity*-based acquisition strategy, we conduct additional experiments where we invert the approach and choose the *least* similar examples from the pool to form the prompt. This investigation aims to ascertain whether the remarkable performance gains can be attributed solely to the semantic similarity between the demonstrations and the test input. The results depicted in Figure 6 substantiate our hypothesis, demonstrating a significant performance drop when employing opposite examples from the pool as in-context exemplars. While this pattern is particularly pronounced in the classification tasks, it consistently emerges across different model sizes and task types. Hence, we can assert that *maximizing semantic similarity between the demonstations and the input test sample* is an unequivocally vital attribute for achieving successful in-context learning outcomes with LLMs. Future endeavors in the field of building effective in-context learning frameworks should incorporate this principle to enable data-efficient algorithms that can fully harness the potential of LLMs.

5.4 Most vs. Least Uncertain Demonstrations

Along these lines, we also opt to examine the duality between selecting the most or the least uncertain in-context examples from the pool. We show the results of these experiments for the GPT models in Figure 7. Interestingly, we observe that while the smaller language models (`gpt2`, `gpt2-medium`, `gpt-large`) perform better with the least uncertain prompts, the larger models seem to start benefiting from the demonstrations with high uncertainty. This is particularly clear in the largest model of our collection, `GPT-Neox` (20B parameters). This interesting finding shows that even larger models will most likely perform better with high entropy in-context examples, similar to their supervised learn-

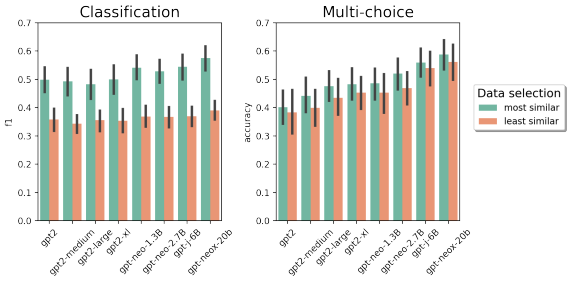


Figure 6: Most vs. least similar in-context examples.

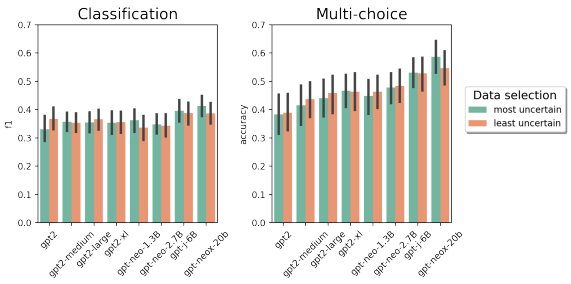


Figure 7: Most vs. least uncertain in-context examples.

ing counterparts. Such findings open a plethora of research questions regarding understanding how in-context learning works (Reynolds and McDonell, 2021; Razeghi et al., 2022; Xie et al., 2022; Min et al., 2022), how AL and data acquisition methods reshape with larger language models or whether we can properly investigate potential emergent abilities of LLMs acquired by model scaling (Wei et al., 2022; Schaeffer et al., 2023).

5.5 Evaluation with Different Metrics

Finally, we want to provide a clear overview of our experiments and summary of our findings, while making some clarifications regarding how we evaluate and compare different approaches to in-context learning. Figure 8 shows the results for in-context learning with random sampling, three data selection techniques inspired by AL (§2.2), namely `diversity`, `uncertainty` and `similarity`, and a zero-shot baseline where no labeled examples are included in the prompt (`no_demo`). We show that in-context learning with k=16 demonstrations consistently outperform zero-shot learning for an average of 15 classification tasks for `gpt2-large`, `gpt-j` and `gpt-neox`. Next, we observe that the best performing in-context example selection method is by a clear margin `similarity`, followed by `diversity`. This finding corroborates the original hypothesis of AL that, indeed, *not all data is equal* and there exist *more informative* data subsets in the pool that can be used as in-context exemplars. We can see that the `uncertainty` baseline, which is usually top performing in supervised AL, generally underperforms in the few-shot setting. Still, there is some evidence that this could change with even larger and better models (§5.4). Finally, delving into the debate on whether ground truth labels matter or not (Min et al., 2022; Yoo et al., 2022), we show that replacing original with random in-context labels hurt significantly the performance of `similarity`, the best data selection method (§5.2).

We further emphasize the significance of employing a meticulous evaluation framework, particularly in the selection of appropriate metrics. In Figure 8, we illustrate the same classification experiments, but with the F_1 score plotted on the left and accuracy on the right. The use of F_1, the conventional metric for classification tasks, reveals a distinct ranking among the various AL methods, with `similarity` exhibiting the best performance, followed by `diversity`. Conversely, when employing accuracy to compare the methods, `diversity` emerges as the top approach, followed by `similarity` and random selection. This disparity highlights the potential for misconceptions or obscured findings, underscoring the need for caution when evaluating and comparing different methods across various models within the in-context learning framework (Dehghani et al., 2021; Min et al., 2022; Yoo et al., 2022; Tedeschi et al., 2023).

6 Related Work

6.1 Understanding In-Context Learning

Few-shot in-context learning with LLMs has garnered significant attention in recent NLP research. Simply concatenating a few labeled examples to form the prompt for the LLM results in high performance gains, even outperforming fine-tuned models (Brown et al., 2020; Chung et al., 2022; Ouyang et al., 2022; Dong et al., 2022). This has naturally lead to study its effectiveness with multiple few-shot learning benchmarks such as `Crossfit` (Ye et al., 2021) and `BigBench` (Srivastava et al., 2022).

Another active area of research is on understanding how in-context learning works (Xie et al., 2022; Garg et al., 2022; Akyürek et al., 2022; Xie et al., 2022; Pan et al., 2023), and what are its strengths and limitations (Webson and Pavlick, 2022; Jang et al., 2022; Levy et al., 2022; Shi et al., 2022; Agrawal et al., 2022; Wei et al., 2023b; Ye et al., 2023b). Previous work has explored the effec-

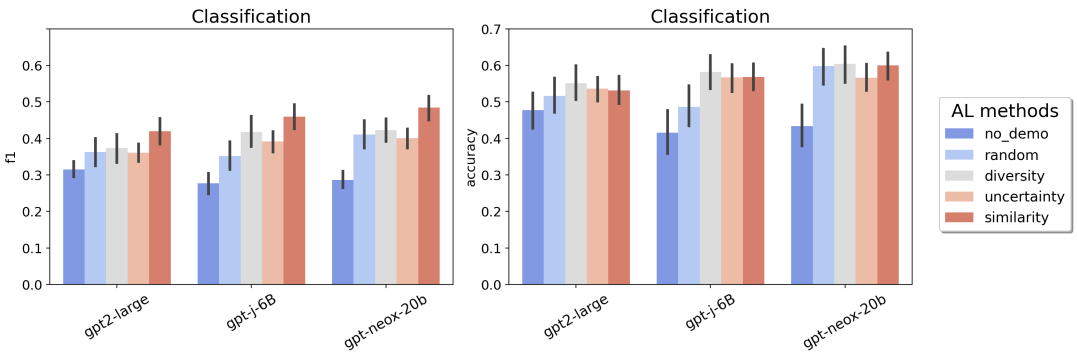


Figure 8: The ranking of data selection methods is different depending on the metric used.

tiveness of the chain-of-thought prompting technique (Wei et al., 2023a; Wang et al., 2022; Madaan and Yazdanbakhsh, 2022), while other studies try to determine the importance of in-context ground truth labels, with Min et al. (2022) showing that random labels do not hurt performance considerably and Yoo et al. (2022) providing a rebuttal. Wei et al. (2023b) explain that model size plays an role in the effect of ground truth labels, showing that small LMs ignore flipped labels, while LLMs can override semantic priors learned during pretraining. Interestingly, Razeghi et al. (2022) demonstrates that in-context learning performance is highly correlated with the prevalence of each instance in the pretraining corpus, showing that models are more accurate on few-shot numerical reasoning on instances whose terms are more frequent.

6.2 Selecting Informative Demonstrations

Typically, work on evaluating LLMs in few-shot settings commonly uses randomly sampled examples to compose the in-context prompt (Brown et al., 2020; Zhang et al., 2022a; Chowdhery et al., 2022; Chung et al., 2022; Touvron et al., 2023). Nonetheless, it has been demonstrated that the effectiveness of few-shot performance significantly depends on the selection of in-context examples (Kocielnik et al., 2022; Ye et al., 2023a; Diao et al., 2023; Xu et al., 2023). Consequently, there is ongoing research on generating or selecting the most informative demonstrations, aiming to maximize the downstream few-shot performance.

Some approaches are based on a retrieval component that sources the most relevant examples from a pool. The prompt retriever can be trainable (Rubin et al., 2022) or based on pretrained embeddings (Liu et al., 2022; Agrawal et al., 2022). Gonen et al. (2022) use uncertainty to evaluate the usefulness of in-context examples and find that the best performing prompts have low perplexity. Zhang et al. (2022b) formulate example selection for in-context learning as a sequential decision problem and show modest performance improvements by acquiring data with their proposed method based on reinforcement learning. Other previous work, instead of focusing on the part of acquiring data for in-context learning, show that demonstration ordering (Lu et al., 2022) and model calibration (Zhao et al., 2021) are additional properties that influence the few-shot learning performance.

6.3 Active Learning for NLP

AL has been extensively studied in various NLP tasks, including machine translation (Miura et al., 2016; Zhao et al., 2020), natural language inference (Snijders et al., 2023), named entity recognition (Shen et al., 2017; Wei et al., 2019), and text classification (Ein-Dor et al., 2020; Margatina et al., 2022; Schröder et al., 2023), among others.

Still, its importance and potential value is on the rise (Zhang et al., 2022c; Rauch et al., 2023), as the current language model pretraining paradigm continues to advance the state-of-the-art (Tamkin et al., 2022). Given the fundamental premise that"not all data is equal" it is reasonable to expect researchers to actively seek the "most informative" data for pretraining or adapting their large language models (LLMs), as well as identifying the most valuable in-context examples for few-shot learning scenarios. Previous work has explored AL for prompt-based finetuning (Köksal et al., 2022), proposing a method based in inter-prompt uncertainty sampling with diversity coupled with the PET architecture (Schick and Schütze, 2021a,b) that outperforms all AL baselines.

7 Conclusion

In this study, we have examined the selection of demonstrations, i.e., labeled data that provide examples of solving a task, for in-context learning with LLMs. We formulated the selection process as a *single iteration active learning problem* and evaluated four standard approaches: `uncertainty`, `diversity`, `similarity`, and `random` sampling. Our evaluation involved 15 models of varying size from the GPT and OPT families, encompassing 15 classification tasks and 9 multi-choice tasks. Through extensive experimentation, we have demonstrated that selecting demonstrations that are semantically similar to the test input examples consistently outperforms all other methods by a significant margin across all model families, sizes, and tasks. This corroborates findings of several previous and concurrent studies that explore the properties of "good" in-context examples (Liu et al., 2022; Shi et al., 2022). Interestingly, our findings reveal that uncertainty sampling, although effective in supervised learning, underperforms in the in-context learning paradigm. This highlights the importance of our work in exploring the principles of active learning in the context of few-shot learning.

Acknowledgements

We would like to thank the anonymous reviewers for their suggestions to improve our work. We also thank Louis Martin, Patrick Lewis, Fabio Petroni and other members of FAIR for their constructive feedback on previous versions of the paper.

Limitations

Tasks & Datasets We acknowledge that even though we experimented with a well established benchmark, the `Crossfit` (Ye et al., 2021) benchmark consisting of 15 classification and 9 multi-choice question answering datasets (Appendix A.1), it might still not be sufficient to ensure that our findings will generalize to any NLP classification or multi-choice application of in-context learning.

Language We also acknowledge that all the datasets and models considered in this work are based on the English language alone. This limits generalizability of our findings to other languages.

Model scale We investigated in-context learning with actively acquired demonstrations with 15 GPT and OPT models that span 125M to 30B parameters. Even though our experimentation is thorough, our findings might not generalize to larger or smaller transformer-based models, or models based in a different architecture.

Active learning considerations We explicitly note in the paper that we do a single active learning iteration, which is different than the common AL loop that consists of multiple iterations. As we explained, because the model-in-the-loop (the LLM) is not updated (no fine-tuning) with new data, performing multiple iterations does not make sense in this context (Figure 2). Still, it would be interesting for future work to explore how we can perform multiple AL iterations while constructing the prompt (i.e., acquiring the demonstrations). The upper bound would be to try all the combinations of a set of labeled data and find the best performing prompt. However, doing this with unlabeled data, in an efficient way, is far from trivial. We refer to Zhang et al. (2022c); Treviso et al. (2023); Margatina and Aletras (2023) for in-depth suggestions for future work in this area.

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2022. In-context examples selection for machine translation.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. 2022. What learning algorithm is in-context learning? investigations with linear models. *ArXiv*, abs/2211.15661.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650.

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136, virtual+Dublin. Association for Computational Linguistics.

Zalán Bodó, Zsolt Minier, and Lehel Csató. 2011. Active learning with clustering. In *Proceedings of the Active Learning and Experimental Design workshop*

In conjunction with AISTATS 2010, volume 16, pages 127–139.

Klaus Brinker. 2003. Incorporating diversity in active learning with support vector machines. In *Proceedings of the International Conference on Machine Learning*, pages 59–66.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4.

Michael Chen, Mike D'Arcy, Alisa Liu, Jared Fernandez, and Doug Downey. 2019. CODAH: An adversarially-authored question answering dataset for common sense. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 63–69, Minneapolis, USA. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. 1996. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4(1):129–145.

Ona de Gibert, Naiara Pérez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20.

Marie-Catherine de Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The commitmentbank: Investigating projection in naturally occurring discourse.

Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. 2021. The benchmark lottery.

Shizhe Diao, Pengcheng Wang, Yong Lin, and Tong Zhang. 2023. Active prompting with chain-of-thought for large language models.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims.

William B Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *ArXiv*, abs/2301.00234.

Liat Ein-Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2020. Active Learning for BERT: An Empirical Study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7949–7962, Online. Association for Computational Linguistics.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep Bayesian active learning with image data. In *Proceedings of the 34th International Conference*

on Machine Learning, volume 70 of *Proceedings of Machine Learning Research*, pages 1183–1192. PMLR.

Xavier García, Yamini Bansal, Colin Cherry, George F. Foster, Maxim Krikun, Fan Feng, Melvin Johnson, and Orhan Firat. 2023. The unreasonable effectiveness of few-shot learning for machine translation.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. 2022. What can transformers learn in-context? a case study of simple function classes. *ArXiv*, abs/2208.01066.

Hila Gonen, Srini Iyer, Terra Blevins, Noah A. Smith, and Luke Zettlemoyer. 2022. Demystifying prompts in language models via perplexity estimation. *ArXiv*, abs/2212.04037.

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 394–398, Montréal, Canada. Association for Computational Linguistics.

Joel Jang, Seonghyeon Ye, and Minjoon Seo. 2022. Can large language models truly understand prompts? a case study with negated prompts. *ArXiv*, abs/2209.12711.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. Qasc: A dataset for question answering via sentence composition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8082–8090.

Hyuhng Joon Kim, Hyunsoo Cho, Junyeob Kim, Taeuk Kim, Kang Min Yoo, and Sang goo Lee. 2022. Self-generated in-context learning: Leveraging autoregressive language models as a demonstration generator. *ArXiv*, abs/2206.08082.

Andreas Kirsch, Tom Rainforth, and Yarin Gal. 2021. Test distribution-aware active learning: A principled approach against distribution shift and outliers.

Rafal Kocielnik, Sara Kangaslahti, Shrimai Prabhumoye, M Hari, R. Michael Alvarez, and Anima Anandkumar. 2022. Can you label less by using out-of-domain data? active & transfer learning with few-shot instructions. *ArXiv*, abs/2211.11798.

Abdullatif Köksal, Timo Schick, and Hinrich Schutze. 2022. Meal: Stable and active learning for few-shot prompting. *ArXiv*, abs/2211.08358.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*.

Itay Levy, Ben Bogin, and Jonathan Berant. 2022. Diverse demonstrations improve in-context compositional generalization.

David D. Lewis and William A. Gale. 1994. A sequential algorithm for training text classifiers. In *In Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

S. Longpre, Julia Reisler, Edward Greg Huang, Yi Lu, Andrew J. Frank, Nikhil Ramesh, and Chris DuBois. 2022. Active learning over multiple domains in natural language tasks. *ArXiv*, abs/2202.00254.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Aman Madaan and Amir Yazdanbakhsh. 2022. Text and patterns: For effective chain of thought, it takes two to tango. *ArXiv*, abs/2209.07686.

Katerina Margatina and Nikolaos Aletras. 2023. On the limitations of simulating active learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4402–4419.

Katerina Margatina, Loic Barrault, and Nikolaos Aletras. 2022. On the importance of effectively adapting pretrained language models for active learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 825–836, Dublin, Ireland. Association for Computational Linguistics.

Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. 2021. Active learning by acquiring contrastive examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 650–663, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Clara H. McCreery, Namit Katariya, Anitha Kannan, Manish Chablani, and Xavier Amatriain. 2020. Effective transfer learning for identifying similar questions: Matching user questions to covid-19 faqs.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work?

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep learning–based text classification: a comprehensive review. *ACM computing surveys (CSUR)*, 54(3):1–40.

Akiva Miura, Graham Neubig, Michael Paul, and Satoshi Nakamura. 2016. Selecting syntactic, non-redundant segments in active learning for machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 20–29, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2022. ETHOS: a multi-label hate speech detection dataset. *Complex Intelligent Systems*, 8(6):4663–4678.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Jane Pan, Tianyu Gao, Howard Chen, and Danqi Chen. 2023. What in-context learning "learns" in-context: Disentangling task recognition and task learning.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2022. Scaling language models: Methods, analysis & insights from training gopher.

Lukas Rauch, Matthias Aßenmacher, Denis Huseljic, Moritz Wirth, Bernd Bischl, and Bernhard Sick. 2023. Activeglae: A benchmark for deep active learning with transformers.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022. Impact of pretraining term frequencies on few-shot numerical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 840–854, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA. Association for Computing Machinery.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, Seattle, United States. Association for Computational Linguistics.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage?

Timo Schick and Hinrich Schütze. 2021a. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021b. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Christopher Schröder, Lydia Müller, Andreas Niekler, and Martin Potthast. 2023. Small-text: Active learning for text classification in python. In *Proceedings of the 17th Conference of the European Chapter of*

the Association for Computational Linguistics: System Demonstrations, pages 84–95, Dubrovnik, Croatia. Association for Computational Linguistics.

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*.

Burr Settles. 2009. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Yanyao Shen, Hyokun Yun, Zachary Lipton, Yakov Kronrod, and Animashree Anandkumar. 2017. Deep active learning for named entity recognition. In *Proceedings of the Workshop on Representation Learning for NLP*, pages 252–256.

Emily Sheng and David C Uthus. 2020. Investigating societal biases in a poetry composition system. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 93–106.

Weijia Shi, Xiaochuang Han, Hila Gonen, Ari Holtzman, Yulia Tsvetkov, and Luke Zettlemoyer. 2022. Toward human readable prompt tuning: Kubrick's the shining is a good movie, and a good prompt too? *ArXiv*, abs/2212.10539.

Ard Snijders, Douwe Kiela, and Katerina Margatina. 2023. Investigating multi-source active learning for natural language inference. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2187–2209, Dubrovnik, Croatia. Association for Computational Linguistics.

Taylor Sorensen, Joshua Robinson, Christopher Rytting, Alexander Shaw, Kyle Rogers, Alexia Delorey, Mahmoud Khalil, Nancy Fulda, and David Wingate. 2022. An information-theoretic approach to prompt engineering without ground truth labels. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 819–862, Dublin, Ireland. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn,

Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramón Risco Delgado, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Timothy Telleen-Lawton, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.

Oyvind Tafjord, Peter Clark, Matt Gardner, Wen-tau Yih, and Ashish Sabharwal. 2019a. Quarel: A dataset and models for answering questions about qualitative relationships. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7063–7071.

Oyvind Tafjord, Matt Gardner, Kevin Lin, and Peter Clark. 2019b. Quartz: An open-domain dataset of qualitative relationship questions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5941–5946.

Alex Tamkin, Dat Pham Nguyen, Salil Deshpande, Jesse Mu, and Noah Goodman. 2022. Active learning helps pretrained models learn the intended task. In *Advances in Neural Information Processing Systems*.

Simone Tedeschi, Johan Bos, Thierry Declerck, Jan Hajic, Daniel Hershcovich, Eduard H. Hovy, Alexander Koller, Simon Krek, Steven Schockaert, Rico Sennrich, Ekaterina Shutova, and Roberto Navigli. 2023. What's the meaning of superhuman performance in today's nlu? *ArXiv*, abs/2305.08414.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Marcos Treviso, Ji-Ung Lee, Tianchu Ji, Betty van Aken, Qingqing Cao, Manuel R. Ciosici, Michael Hassid, Kenneth Heafield, Sara Hooker, Colin Raffel, Pedro H. Martins, André F. T. Martins, Jessica Zosa Forde, Peter Milder, Edwin Simpson, Noam Slonim, Jesse Dodge, Emma Strubell, Niranjan Balasubramanian, Leon Derczynski, Iryna Gurevych, and Roy Schwartz. 2023. Efficient Methods for Natural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics*, 11:826–860.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2022. Towards understanding chain-of-thought prompting: An empirical study of what matters. *ArXiv*, abs/2212.10001.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of*

the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2300–2344, Seattle, United States. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023a. Chain-of-thought prompting elicits reasoning in large language models.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. 2023b. Larger language models do in-context learning differently.

Qiang Wei, Yukun Chen, Mandana Salimi, Joshua C Denny, Qiaozhu Mei, Thomas A Lasko, Qingxia Chen, Stephen Wu, Amy Franklin, Trevor Cohen, and Hua Xu. 2019. Cost-aware active learning for named entity recognition in clinical text. *Journal of the American Medical Informatics Association*, 26(11):1314–1322.

Zhiyong Wu, Yaoxiang Wang, Jiacheng Ye, and Lingpeng Kong. 2022. Self-adaptive in-context learning. *ArXiv*, abs/2212.10375.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*.

Canwen Xu, Yichong Xu, Shuohang Wang, Yang Liu, Chenguang Zhu, and Julian McAuley. 2023. Small models are valuable plug-ins for large language models.

Sohee Yang, Jonghyeon Kim, Joel Jang, Seonghyeon Ye, Hyunji Lee, and Minjoon Seo. 2023. Improving probability-based prompt selection through unified evaluation and analysis.

Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. 2023a. Compositional exemplars for in-context learning.

Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. 2021. CrossFit: A few-shot learning challenge for cross-task generalization in NLP. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7163–7189, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xi Ye, Srinivasan Iyer, Asli Celikyilmaz, Ves Stoyanov, Greg Durrett, and Ramakanth Pasunuru. 2023b. Complementary explanations for effective in-context learning. In *Findings of the Conference of the Association for Computational Linguistics*.

Kang Min Yoo, Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang-goo Lee, and Taeuk Kim. 2022. Ground-truth labels matter: A deeper look into input-label demonstrations.

W. Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2022. Generate rather than retrieve: Large language models are strong context generators. *ArXiv*, abs/2209.10063.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022a. Opt: Open pre-trained transformer language models.

Yiming Zhang, Shi Feng, and Chenhao Tan. 2022b. Active example selection for in-context learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9134–9148, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhisong Zhang, Emma Strubell, and Eduard Hovy. 2022c. A survey of active learning for natural language processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Tony Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. *ICML*, abs/2102.09690.

Yuekai Zhao, Haoran Zhang, Shuchang Zhou, and Zhihua Zhang. 2020. Active learning approaches to enhancing neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1796–1806, Online. Association for Computational Linguistics.

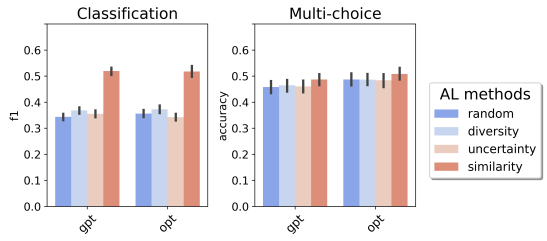


Figure 9: Results per model family.

A Experimental Details

A.1 Tasks & Datasets

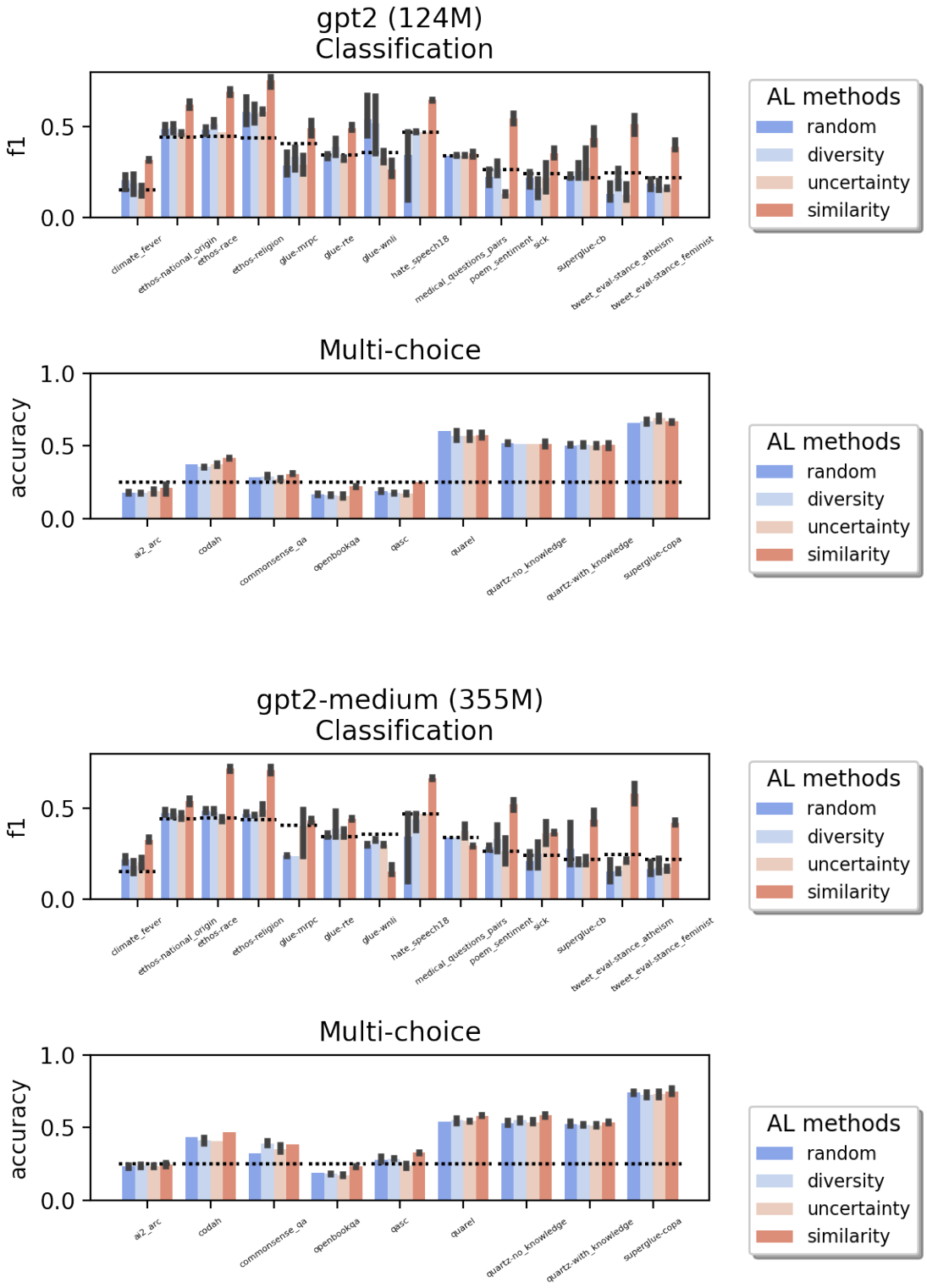
Following Min et al. (2022), we evaluate our models in 15 classification and 9 multi-choice tasks taken from the `Crossfit` (Ye et al., 2021) benchmark. Specifically the tasks we evaluate are *poem_sentiment* (Sheng and Uthus, 2020), *glue-wnli* (Wang et al., 2019; Levesque et al., 2012), *climate_fever* (Diggelmann et al., 2020), *glue-rte* (Wang et al., 2019), *superglue-cb* (de Marneffe et al., 2019), *sick* (Minaee et al., 2021), *medical_questions_pairs* (McCreery et al., 2020), *glue-mrpc* (Wang et al., 2019; Dolan and Brockett, 2005), *hate_speech18* (de Gibert et al., 2018), *ethos-national_origin* (Mollas et al., 2022), *ethos-race* (Mollas et al., 2022), *ethos-religion* (Mollas et al., 2022), *tweet_eval-stance_atheism* (Barbieri et al., 2020), *tweet_eval-stance_feminist* (Barbieri et al., 2020) and *quarel* (Tafjord et al., 2019a), *openbookqa*,*qasc* (Khot et al., 2020), *commonsense_qa*, *ai2_arc* (Clark et al., 2018), *codah* (Chen et al., 2019), *superglue-copa* (Gordon et al., 2012), *quartz-with_knowledge* (Tafjord et al., 2019b), *quartz-no_knowledge* (Tafjord et al., 2019b), for classification and multi-choice respectively.

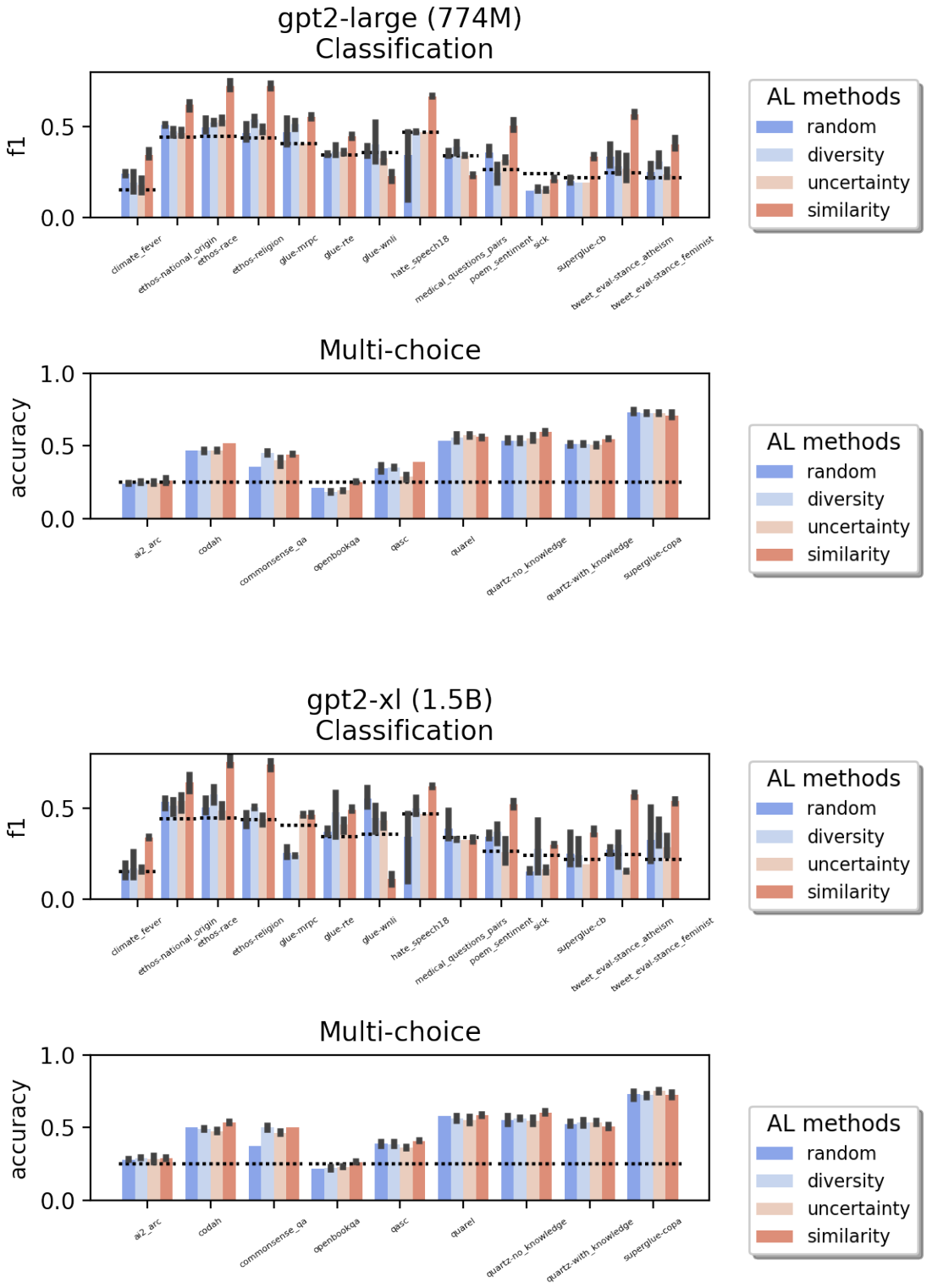
A.2 Full results

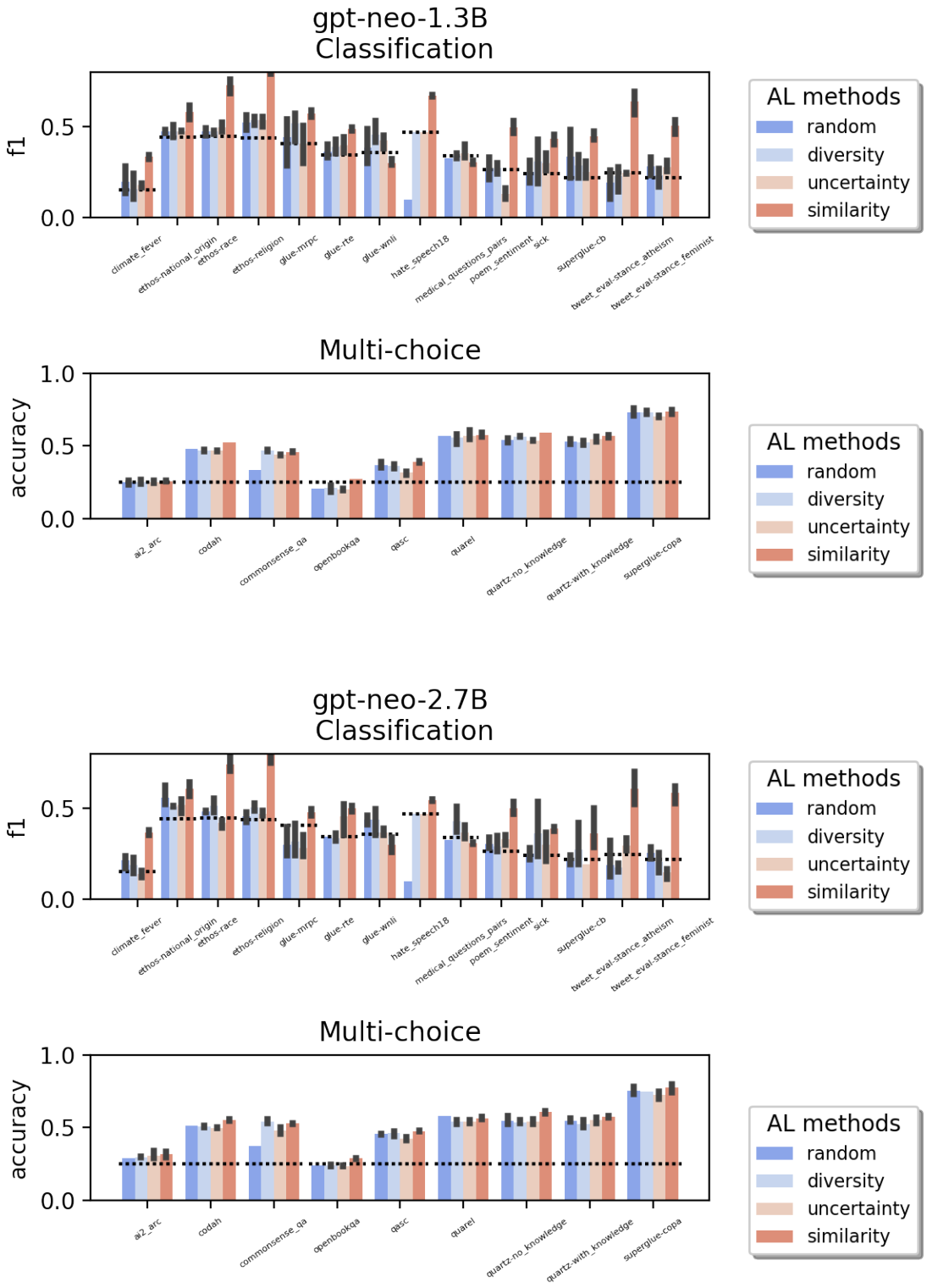
We provide below the full set of results, for each dataset, model and active learning acquisition strategy considered. The dashed line depicts the majority vote baseline.

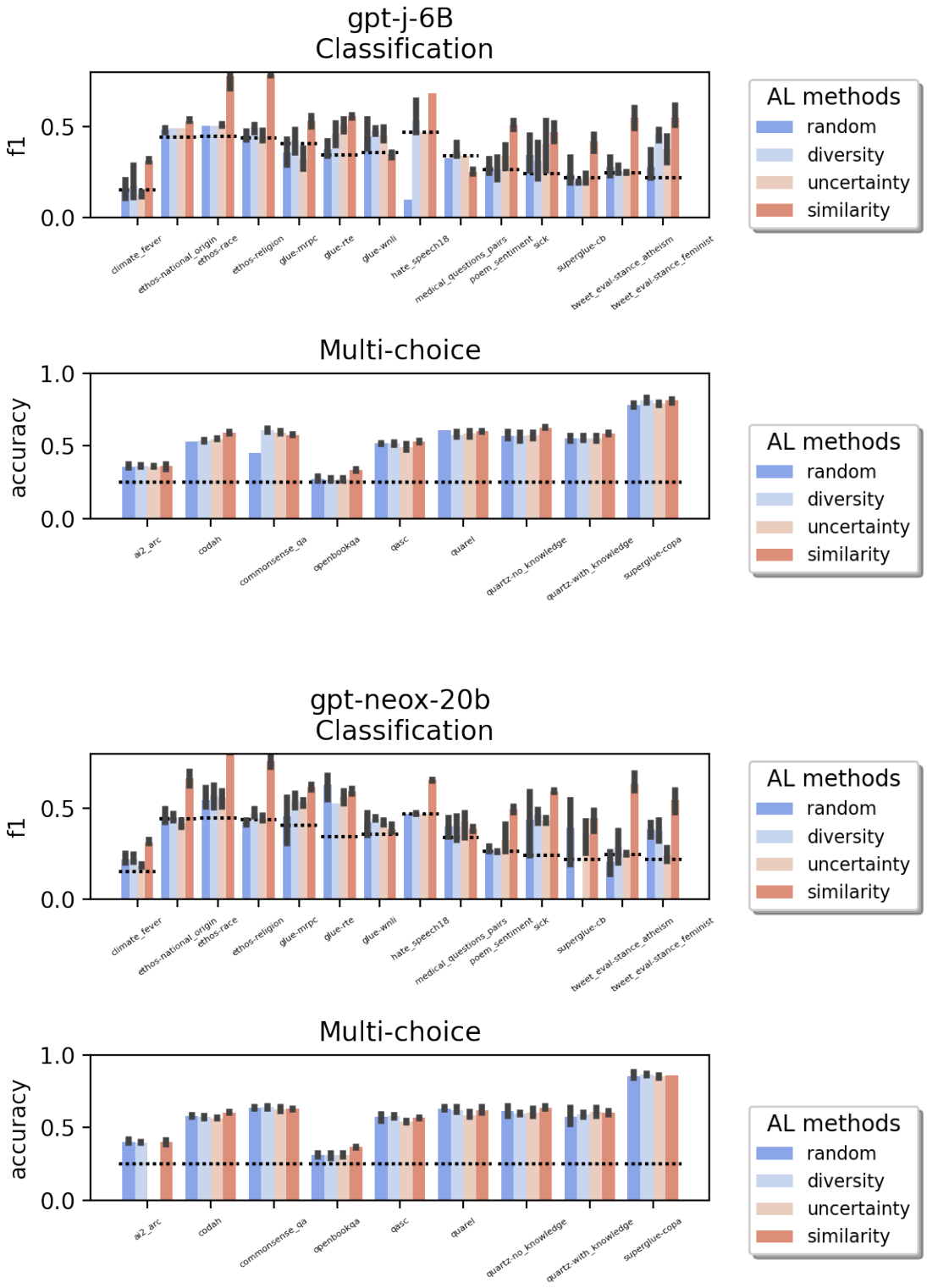
A.3 Model Family

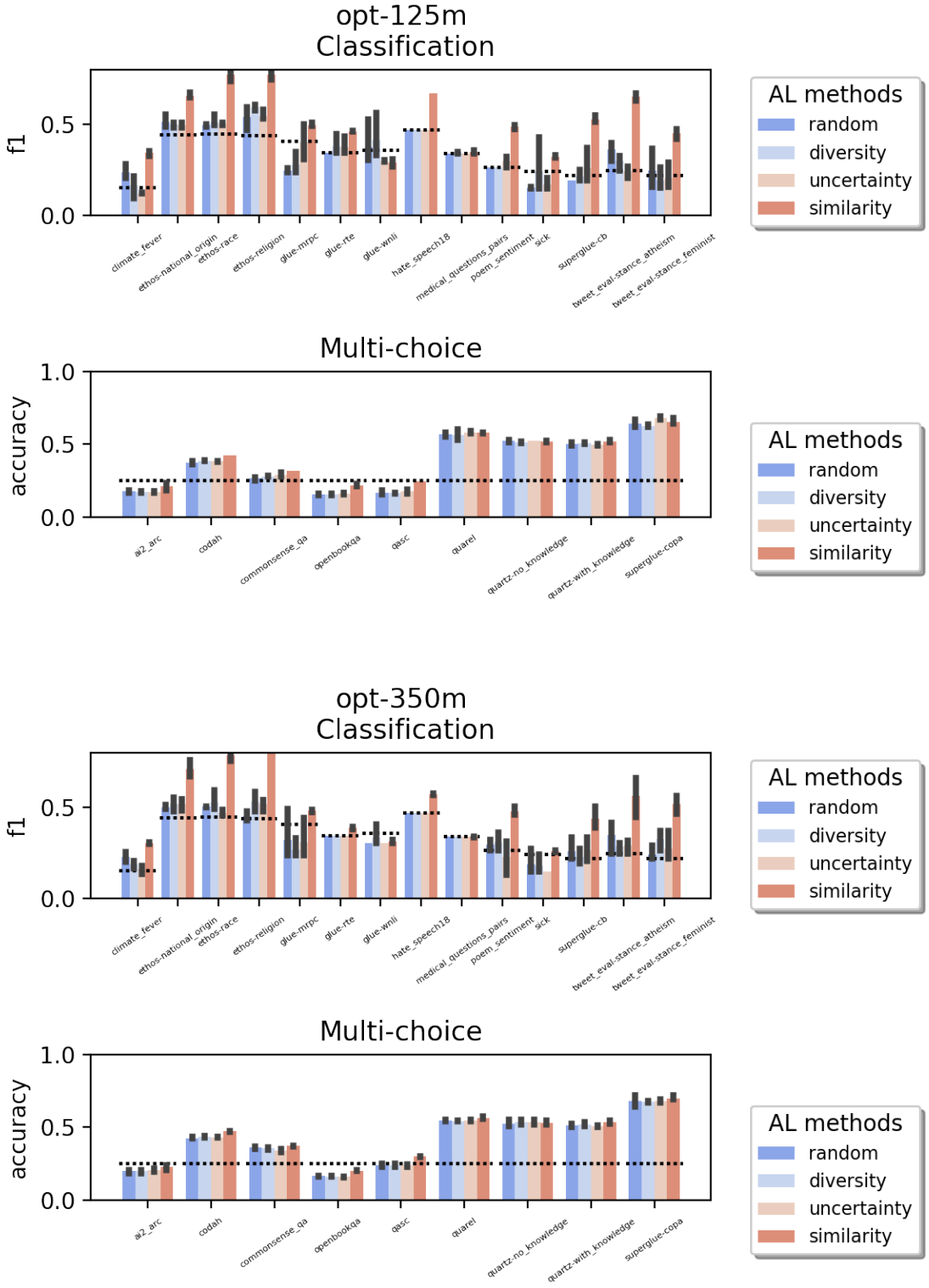
We provide the results on few-shot learning with k=16 demonstrations per prompt per model family and task type in Figure 9. We observe the same patterns for both GPT and OPT models.

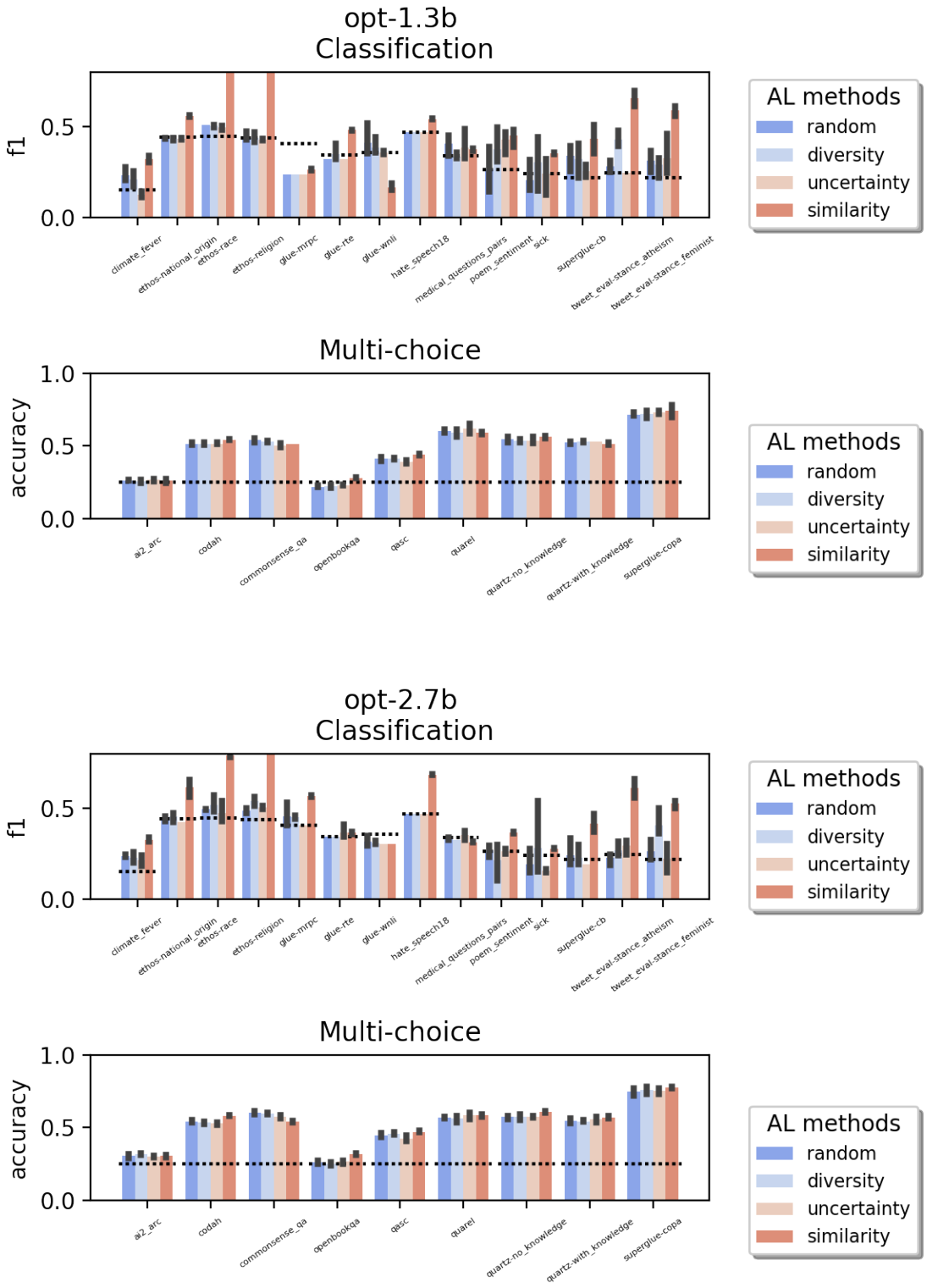

gpt2 (124M)
Classification
f1
0.0
0.5
climate_fever
ethos-national_origin
ethos-race
ethos-religion
glue-mrpc
glue-rte
glue-wnli
hate_speech18
medical_questions_pairs
poem_sentiment
sick
superglue-cb
tweet_eval-stance_atheism
tweet_eval-stance_feminist
AL methods
random
diversity
uncertainty
similarity
Multi-choice
accuracy
0.0
0.5
1.0
ai2_arc
codah
commonsense_qa
openbookqa
qasc
quarel
quartz-no_knowledge
quartz-with_knowledge
superglue-copa
AL methods
random
diversity
uncertainty
similarity
gpt2-medium (355M)
Classification
f1
0.0
0.5
climate_fever
ethos-national_origin
ethos-race
ethos-religion
glue-mrpc
glue-rte
glue-wnli
hate_speech18
medical_questions_pairs
poem_sentiment
sick
superglue-cb
tweet_eval-stance_atheism
tweet_eval-stance_feminist
AL methods
random
diversity
uncertainty
similarity
Multi-choice
accuracy
0.0
0.5
1.0
ai2_arc
codah
commonsense_qa
openbookqa
qasc
quarel
quartz-no_knowledge
quartz-with_knowledge
superglue-copa
AL methods
random
diversity
uncertainty
similarity

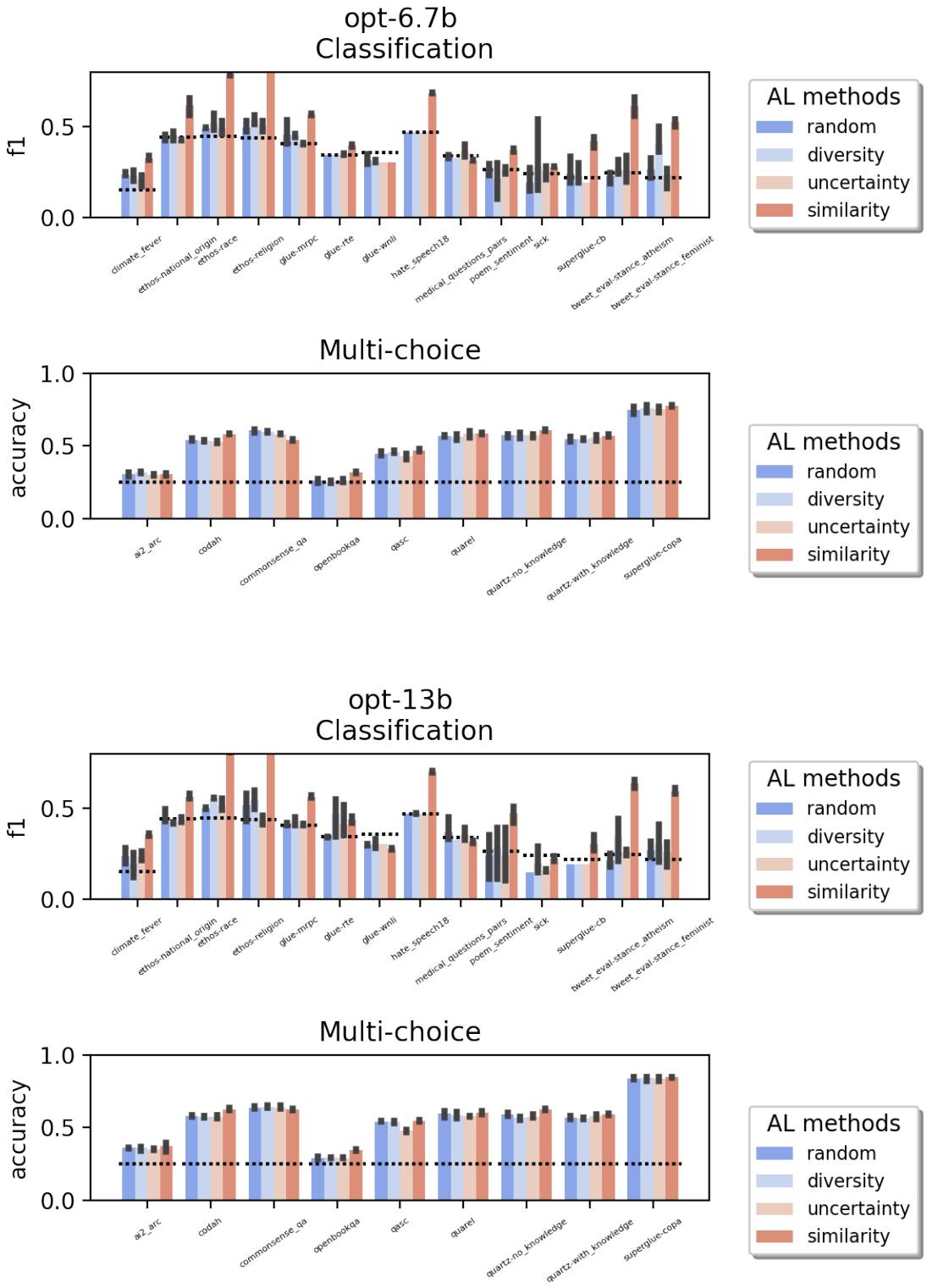

gpt2-large (774M)
Classification
f1
0.0
0.5
climate_fever
ethos-national_origin
ethos-race
ethos-religion
glue-mrpc
glue-rte
glue-wnli
hate_speech18
medical_questions_pairs
poem_sentiment
sick
superglue-cb
tweet_eval-stance_atheism
tweet_eval-stance_feminist
AL methods
random
diversity
uncertainty
similarity
Multi-choice
accuracy
0.0
0.5
1.0
ai2_arc
codah
commonsense_qa
openbookqa
qasc
quarel
quartz-no_knowledge
quartz-with_knowledge
superglue-copa
AL methods
random
diversity
uncertainty
similarity
gpt2-xl (1.5B)
Classification
f1
0.0
0.5
climate_fever
ethos-national_origin
ethos-race
ethos-religion
glue-mrpc
glue-rte
glue-wnli
hate_speech18
medical_questions_pairs
poem_sentiment
sick
superglue-cb
tweet_eval-stance_atheism
tweet_eval-stance_feminist
AL methods
random
diversity
uncertainty
similarity
Multi-choice
accuracy
0.0
0.5
1.0
ai2_arc
codah
commonsense_qa
openbookqa
qasc
quarel
quartz-no_knowledge
quartz-with_knowledge
superglue-copa
AL methods
random
diversity
uncertainty
similarity

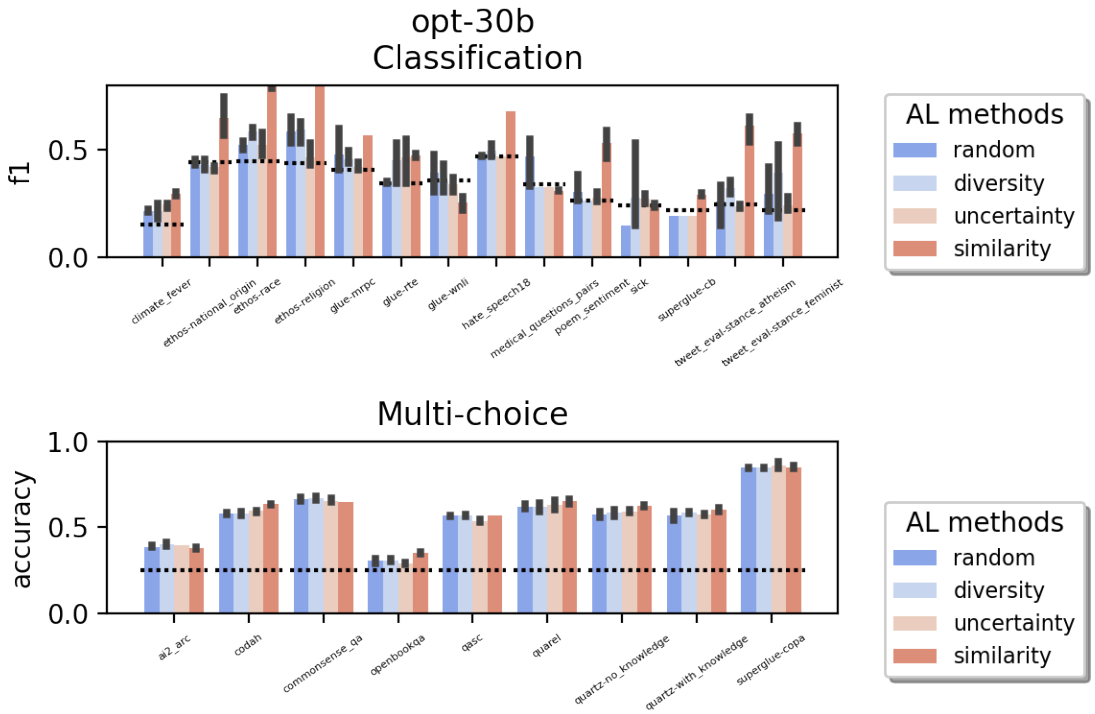

gpt-neo-1.3B
Classification
f1
0.0
0.5
climate_fever
ethos-national_origin
ethos-race
ethos-religion
glue-mrpc
glue-rte
glue-wnli
hate_speech18
medical_questions_pairs
poem_sentiment
sick
superglue-cb
tweet_eval-stance_atheism
tweet_eval-stance_feminist
AL methods
random
diversity
uncertainty
similarity
Multi-choice
accuracy
0.0
0.5
1.0
ai2_arc
codah
commonsense_qa
openbookqa
qasc
quarel
quartz-no_knowledge
quartz-with_knowledge
superglue-copa
AL methods
random
diversity
uncertainty
similarity
gpt-neo-2.7B
Classification
f1
0.0
0.5
climate_fever
ethos-national_origin
ethos-race
ethos-religion
glue-mrpc
glue-rte
glue-wnli
hate_speech18
medical_questions_pairs
poem_sentiment
sick
superglue-cb
tweet_eval-stance_atheism
tweet_eval-stance_feminist
AL methods
random
diversity
uncertainty
similarity
Multi-choice
accuracy
0.0
0.5
1.0
ai2_arc
codah
commonsense_qa
openbookqa
qasc
quarel
quartz-no_knowledge
quartz-with_knowledge
superglue-copa
AL methods
random
diversity
uncertainty
similarity


gpt-j-6B
Classification
f1
0.0
0.5
climate_fever
ethos-national_origin
ethos-race
ethos-religion
glue-mrpc
glue-rte
glue-wnli
hate_speech18
medical_questions_pairs
poem_sentiment
sick
superglue-cb
tweet_eval-stance_atheism
tweet_eval-stance_feminist
AL methods
random
diversity
uncertainty
similarity
Multi-choice
accuracy
0.0
0.5
1.0
ai2_arc
codah
commonsense_qa
openbookqa
qasc
quarel
quartz-no_knowledge
quartz-with_knowledge
superglue-copa
AL methods
random
diversity
uncertainty
similarity
gpt-neox-20b
Classification
f1
0.0
0.5
climate_fever
ethos-national_origin
ethos-race
ethos-religion
glue-mrpc
glue-rte
glue-wnli
hate_speech18
medical_questions_pairs
poem_sentiment
sick
superglue-cb
tweet_eval-stance_atheism
tweet_eval-stance_feminist
AL methods
random
diversity
uncertainty
similarity
Multi-choice
accuracy
0.0
0.5
1.0
ai2_arc
codah
commonsense_qa
openbookqa
qasc
quarel
quartz-no_knowledge
quartz-with_knowledge
superglue-copa
AL methods
random
diversity
uncertainty
similarity


opt-125m
Classification
f1
0.0
0.5
climate_fever
ethos-national_origin
ethos-race
ethos-religion
glue-mrpc
glue-rte
glue-wnli
hate_speech18
medical_questions_pairs
poem_sentiment
sick
superglue-cb
tweet_eval-stance_atheism
tweet_eval-stance_feminist
AL methods
random
diversity
uncertainty
similarity
Multi-choice
accuracy
0.0
0.5
1.0
ai2_arc
codah
commonsense_qa
openbookqa
qasc
quarel
quartz-no_knowledge
quartz-with_knowledge
superglue-copa
AL methods
random
diversity
uncertainty
similarity
opt-350m
Classification
f1
0.0
0.5
climate_fever
ethos-national_origin
ethos-race
ethos-religion
glue-mrpc
glue-rte
glue-wnli
hate_speech18
medical_questions_pairs
poem_sentiment
sick
superglue-cb
tweet_eval-stance_atheism
tweet_eval-stance_feminist
AL methods
random
diversity
uncertainty
similarity
Multi-choice
accuracy
0.0
0.5
1.0
ai2_arc
codah
commonsense_qa
openbookqa
qasc
quarel
quartz-no_knowledge
quartz-with_knowledge
superglue-copa
AL methods
random
diversity
uncertainty
similarity


opt-1.3b
Classification
f1
0.0
0.5
climate_fever
ethos-national_origin
ethos-race
ethos-religion
glue-mrpc
glue-rte
glue-wnli
hate_speech18
medical_questions_pairs
poem_sentiment
sick
superglue-cb
tweet_eval-stance_atheism
tweet_eval-stance_feminist
AL methods
random
diversity
uncertainty
similarity
Multi-choice
accuracy
0.0
0.5
1.0
ai2_arc
codah
commonsense_qa
openbookqa
qasc
quarel
quartz-no_knowledge
quartz-with_knowledge
superglue-copa
AL methods
random
diversity
uncertainty
similarity
opt-2.7b
Classification
f1
0.0
0.5
climate_fever
ethos-national_origin
ethos-race
ethos-religion
glue-mrpc
glue-rte
glue-wnli
hate_speech18
medical_questions_pairs
poem_sentiment
sick
superglue-cb
tweet_eval-stance_atheism
tweet_eval-stance_feminist
AL methods
random
diversity
uncertainty
similarity
Multi-choice
accuracy
0.0
0.5
1.0
ai2_arc
codah
commonsense_qa
openbookqa
qasc
quarel
quartz-no_knowledge
quartz-with_knowledge
superglue-copa
AL methods
random
diversity
uncertainty
similarity


opt-6.7b
Classification
f1
0.0
0.5
climate_fever
ethos-national_origin
ethos-race
ethos-religion
glue-mrpc
glue-rte
glue-wnli
hate_speech18
medical_questions_pairs
poem_sentiment
sick
superglue-cb
tweet_eval-stance_atheism
tweet_eval-stance_feminist
AL methods
random
diversity
uncertainty
similarity
Multi-choice
accuracy
0.0
0.5
1.0
ai2_arc
codah
commonsense_qa
openbookqa
qasc
quarel
quartz-no_knowledge
quartz-with_knowledge
superglue-copa
AL methods
random
diversity
uncertainty
similarity
opt-13b
Classification
f1
0.0
0.5
climate_fever
ethos-national_origin
ethos-race
ethos-religion
glue-mrpc
glue-rte
glue-wnli
hate_speech18
medical_questions_pairs
poem_sentiment
sick
superglue-cb
tweet_eval-stance_atheism
tweet_eval-stance_feminist
AL methods
random
diversity
uncertainty
similarity
Multi-choice
accuracy
0.0
0.5
1.0
ai2_arc
codah
commonsense_qa
openbookqa
qasc
quarel
quartz-no_knowledge
quartz-with_knowledge
superglue-copa
AL methods
random
diversity
uncertainty
similarity


opt-30b
Classification
f1
0.0
0.5
climate_fever
ethos-national_origin
ethos-race
ethos-religion
glue-mrpc
glue-rte
glue-wnli
hate_speech18
medical_questions_pairs
poem_sentiment
sick
superglue-cb
tweet_eval-stance_atheism
tweet_eval-stance_feminist
AL methods
random
diversity
uncertainty
similarity
Multi-choice
accuracy
0.0
0.5
1.0
ai2_arc
codah
commonsense_qa
openbookqa
qasc
quarel
quartz-no_knowledge
quartz-with_knowledge
superglue-copa
AL methods
random
diversity
uncertainty
similarity

4.3 Impact

According to Google Scholar, the paper has received 22 citations as of May 2024. It was featured in various surveys (e.g. Wan et al., 2023; Zheng et al., 2023; Li et al., 2024; Tan et al., 2024) and has influenced follow-up work (Delaflor et al., 2022; Bansal and Sharma, 2023; Gao et al., 2023; Muldrew et al., 2024; Pang et al., 2024; Wang, 2024; Rouzegar and Makrehchi, 2024).

4.4 Discussion

In this study, we have presented a comprehensive analysis of data selection methods for in-context learning with Language Model Models (LLMs). Our findings highlight several key takeaways and shed light on important limitations in the current understanding of this emerging field. Notably, our results demonstrate the paramount importance of selecting in-context examples that closely align with the semantic content of the input test examples. This finding is consistent across model families, sizes, and tasks, underscoring the significance of semantic similarity in driving successful in-context learning outcomes. Furthermore, our study challenges the conventional idea regarding the efficacy of uncertainty sampling, revealing its underperformance in the in-context learning paradigm compared to traditional supervised learning settings. This prompts a reevaluation of active learning principles in the context of few-shot learning with LLMs, emphasizing the need for tailored approaches to data selection in this domain.

Despite these insightful findings, our study has several limitations that warrant consideration. While our analysis encompasses a diverse range of model sizes from both GPT and OPT families, the scalability of our findings to even larger or smaller transformer-based models remains an open question. Furthermore, our study focuses on a single iteration of active learning, which may not fully capture the potential benefits of iterative data selection strategies. Exploring multi-iteration active learning frameworks in the context of in-context learning with LLMs could be a promising avenue for future research, albeit challenging in practice.

Chapter 5

Publication IV: Limitations of Simulating Active Learning

The main contribution of this chapter is the paper *On the Limitations of Simulating Active Learning*, which was published at the *Findings of the Association for Computational Linguistics* at the ACL conference in July 2023. We first outline the motivation for this work (Section 5.1), followed by the paper itself (Section 5.2), the impact that it has had so far (Section 5.3), and discussion (Section 5.4).

5.1 Introduction

Throughout this PhD thesis, we delved deeply into the realm of active learning for natural language processing, conducting exhaustive research on numerous algorithms across various stages of the pipeline and with diverse settings and models. Crucially, we not only studied these algorithms but also meticulously implemented them, conducting thousands of experiments for our previously published papers. These endeavors have yielded a wealth of insights, hard-earned lessons, key findings, and fertile ideas for future exploration. As we amassed a trove of questions, unresolved issues, insights, and practical guidelines, we recognized the value in consolidating them into a comprehensive position paper for peer review and dissemination within the community. Thus, our motivation for the final paper of this PhD thesis emerged—a synthesis of our years of exploration into active learning algorithms for Language Models (LMs), encapsulating our collective knowledge and experiences gained throughout this journey.

5.2 The Paper

Author Contributions

The paper is co-authored by Nikolaos Aletras and myself. As a lead author, I collected the material for the position paper and wrote it. We had multiple discussions with Nikolaos and he helped revise the final version of the paper.

On the Limitations of *Simulating* Active Learning

Katerina Margatina **Nikolaos Aletras**
University of Sheffield
{k.margatina, n.aletras}@sheffield.ac.uk

Abstract

Active learning (AL) is a *human-and-model-in-the-loop* paradigm that iteratively selects informative unlabeled data for human annotation, aiming to improve over random sampling. However, performing AL experiments with human annotations on-the-fly is a laborious and expensive process, thus unrealistic for academic research. An easy fix to this impediment is to *simulate* AL, by treating an *already* labeled and publicly available dataset as the pool of *unlabeled* data. In this position paper, we first survey recent literature and highlight the challenges across all different steps within the AL loop. We further unveil neglected caveats in the experimental setup that can significantly affect the quality of AL research. We continue with an exploration of how the *simulation* setting can govern empirical findings, arguing that it might be one of the answers behind the ever posed question *"why do active learning algorithms sometimes fail to outperform random sampling?*". We argue that evaluating AL algorithms on available labeled datasets might provide a *lower bound* as to their effectiveness in real data. We believe it is essential to collectively shape the best practices for AL research, particularly as engineering advancements in LLMs push the research focus towards data-driven approaches (e.g., data efficiency, alignment, fairness). In light of this, we have developed guidelines for future work. Our aim is to draw attention to these limitations within the community, in the hope of finding ways to address them.

1 Introduction

Based on the assumption that "*not all data is equal*", active learning (AL) (Cohn et al., 1996; Settles, 2009) aims to identify the most informative data for annotation from a pool (or a stream) of unlabeled data (i.e., data acquisition). With multiple rounds of model training, data acquisition and human annotation (Figure 1), the goal is to achieve

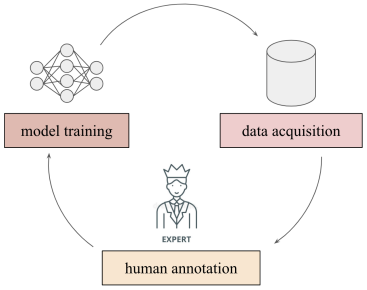


Figure 1: High-level overview of the *train-acquire-annotate* steps of the active learning loop.

data efficiency. A data efficient AL algorithm entails that a model achieves satisfactory performance on a held-out test set, by being trained with only a fraction of the acquired data.

AL has traditionally attracted wide attention in the Natural Language Processing (NLP) community. It has been explored for machine translation (Haffari et al., 2009; Dara et al., 2014; Miura et al., 2016; Zhao et al., 2020), text classification (Ein-Dor et al., 2020; Schröder and Niekler, 2020; Margatina et al., 2022; Schröder et al., 2023), part-of-speech tagging (Chaudhary et al., 2021), coreference (Yuan et al., 2022) and entity resolution (Qian et al., 2017; Kasai et al., 2019), named entity recognition (Erdmann et al., 2019; Shen et al., 2017; Wei et al., 2019; Shelmanov et al., 2021), and natural language inference (Snijders et al., 2023), *inter alia*. However, its potential value is still growing (Zhang et al., 2022d), driven by advancements in the state-of-the-art in language model pretraining (Tamkin et al., 2022). Given the initial assumption that "not all data is equal", it is reasonable to expect researchers to seek out the "most valuable" data for pretraining or adapting their language models.

The usual pool-based AL setting is to acquire data from an unlabeled pool, label it, and use it to

Findings of the Association for Computational Linguistics: ACL 2023, pages 4402–4419
July 9-14, 2023

train a supervised model that, hopefully, obtains satisfactory performance on a test set for the task at hand. This is very similar to the general model-in-the-loop paradigm (Karmakharm et al., 2019; Bartolo et al., 2020, 2022; Kiela et al., 2021; Wallace et al., 2022), with the main difference being the AL-based data acquisition stage. The assumption is that, by iteratively selecting data for annotation according to an informativeness criterion, it will result into better model predictive performance compared to randomly sampling and annotate data of the same size.

However, this does not always seem to be the case. A body of work has shown that AL algorithms, that make use of uncertainty (Lewis and Gale, 1994; Cohn et al., 1996; Houlsby et al., 2011; Gal et al., 2017), diversity sampling (Brinker, 2003; Bodó et al., 2011; Sener and Savarese, 2018) or even more complex acquisition strategies (Ducoffe and Precioso, 2018; Ash et al., 2020; Yuan et al., 2020; Margatina et al., 2021), often fail to improve over a simple random sampling baseline (Baldridge and Palmer, 2009; Ducoffe and Precioso, 2018; Lowell et al., 2019; Kees et al., 2021; Karamcheti et al., 2021; Snijders et al., 2023). Such findings pose a serious question on the practical usefulness of AL, as they do not corroborate its initial core hypothesis that *not all data is equally useful for training a model*. In other words, if we cannot show that one subset of the data is "better"[1] than another, why do AL in the first place?

Only a small body of work has attempted to explore the pain points of AL. For instance, Karamcheti et al. (2021), leveraging visualisations from *data maps* (Swayamdipta et al., 2020), show that AL algorithms tend to acquire *collective outliers* (i.e. groups of examples that deviate from the rest of the examples but cluster together), thus explaining the utter failure of eight AL algorithms to outperform random sampling in visual question answering. Building on this work, more recently Snijders et al. (2023) corroborate these findings for the task of natural language inference and further show that uncertainty based AL methods recover and even surpass random selection when hard-to-learn data points are removed from the pool. Lowell et al. (2019) show that the benefits of AL with certain models and domains do not generalize reliably across models and tasks. This could be problematic since, in practice, one might not have the means to explore and compare alternative AL strategies. They also show that an actively acquired dataset using a certain model-in-the-loop, may be disadvantageous for training models of a different family, raising the issue of whether the downsides inherent to AL are worth the modest and inconsistent performance gains it tends to afford.

In this paper, we aim to explore all possible limitations that researchers and practitioners currently face when doing research on AL (Zhang et al., 2022d). We first describe the process of pool-based AL (Figure 1) and identify challenges in every step of the iterative process (§2). Next, we unearth obscure details that are often left unstated and under-explored (§3). We then delve into a more philosophical discussion of the role of simulation and its connection to real practical applications (§4). Finally, we provide guidelines for future work (§5) and conclusions (§6), aspiring to promote neglected, but valuable, ideas to improve the direction of research in active learning.

2 Challenges in the Active Learning Loop

We first introduce the typical steps in the pool-based AL setting (Lewis and Gale, 1994) and identify several challenges that an AL practitioner has to deal with, across all steps (Figure 2).[2]

2.1 Problem Definition

Consider the experimental scenario where we want to model a specific NLP task for which we do not yet have any labeled data, but we have access to a large pool of unlabeled data $\mathcal{D}_{\text{pool}}$. We assume that it is unrealistic (e.g., laborious, expensive) to have humans annotating all of it. $\mathcal{D}_{\text{pool}}$ constitutes the textual corpus from which we want to sample a fraction of the most *useful* (e.g., informative, representative) data points for human annotation. In order to perform active learning, we need an initial labeled dataset $\mathcal{D}_{\text{lab}}$, often called "seed" dataset, to be used for training a task-specific model with supervised learning. To evaluate the model, we need a usually small validation set for model selection $\mathcal{D}_{\text{val}}$ and a held out test set $\mathcal{D}_{\text{test}}$ to evaluate the model's generalization. We use $\mathcal{D}_{\text{lab}}$ and $\mathcal{D}_{\text{val}}$ to train the first model and then test it on $\mathcal{D}_{\text{test}}$.

[1] We consider a labeled dataset $A \subset C$ to be "better" than a labeled dataset $B \subset C$, both sampled from a corpus C and $|A| = |B|$, if a model M_A trained on A yields higher performance on a test set compared to M_B, where both models are identical in terms of architecture, training procedure, etc.

[2] We point the reader to the comprehensive survey of Zhang et al. (2022d) for a more in-depth exploration of recent literature in AL.

In this stage, we start acquiring labeled data for model training. Data points are sampled from $\mathcal{D}_{\text{pool}}$ via an acquisition strategy and subsequently passed to human annotators for labeling. The acquisition function selects a batch of data $Q \subset \mathcal{D}_{\text{pool}}$ according to some informativeness criterion and can either use the model-in-the-loop or not. We employ crowdsourcing or expert annotators to label the selected batch Q which then is appended to the labeled dataset $\mathcal{D}_{\text{lab}}$.

Now that we have augmented the seed dataset with more data, we re-train the model on the new training dataset, $\mathcal{D}_{\text{lab}}$. We test the new model on $\mathcal{D}_{\text{test}}$ and we stop if we obtain satisfactory performance or if the budget for annotation has run out (or using any other stopping criterion). If we do not want to stop, we use the acquisition function to select more unlabeled data from $\mathcal{D}_{\text{pool}}$, which we annotate and append to $\mathcal{D}_{\text{lab}}$, etc. This is the AL loop shown in Figure 2.

2.2 Active Learning Design

Seed dataset We start the AL loop (§2.1) by defining an initial labeled "seed dataset" (Figure 2: 1). The seed dataset plays an important role, as it will be used to train the the first model-in-the-loop (Tomanek et al., 2009; Horbach and Palmer, 2016). In AL research, we typically address the cold-start problem by sampling from $\mathcal{D}_{\text{pool}}$ with a uniform distribution for each class, either retaining the true label distribution or choosing data that form a balanced label distribution.[3] This is merely a convenient design choice, as it is simple and easy to implement. However, sampling the seed dataset this way, does not really reflect a real-world setting where the label distribution of the (unlabeled data of the) pool is actually unknown.

Prabhu et al. (2019) performed a study of such sampling bias in AL, showing no effect in different seed datasets across the considered methods. Ein-Dor et al. (2020) also experimented with different imbalanced seed datasets, showing that AL improves over random sampling in settings with highest imbalance.

Furthermore, the choice of the seed dataset has a direct effect on the entire AL design because the first model-in-the-loop marks the reference point of the performance in $\mathcal{D}_{\text{test}}$. In other words, the performance of the first model is essentially the baseline, according to which a practitioner will plan the AL loop based on the goal performance and the available budget. It is thus essential to revisit existing approaches on choosing the seed dataset (Kang et al., 2004; Vlachos, 2006; Hu et al., 2010; Yuan et al., 2020) and evaluate them towards a realistic simulation of an AL experiment.

Number of iterations & acquisition budget After choosing the seed dataset it is natural to decide the number of iterations, the acquisition size (the size of the acquired batch $\mathcal{Q}$) and the budget (the size of the actively collected $\mathcal{D}_{\text{lab}}$) of the AL experiment. This is another part where literature does not offer concrete explanations on the design choice. Papers that address the cold-start problem would naturally focus on the very few first AL iterations (Yuan et al., 2020), while others might simulate AL until a certain percentage of the pool has been annotated (Prabhu et al., 2019; Lowell et al., 2019; Zhao et al., 2020; Zhang and Plank, 2021; Margatina et al., 2022) or until a certain fixed and predefined number of examples has been annotated (Ein-Dor et al., 2020; Kirsch et al., 2021).

2.3 Model Training

We now train the model-in-the-loop with the available labeled dataset $\mathcal{D}_{\text{lab}}$ (Figure 2: 2). Interestingly, there are not many studies that explore how we should properly train the model in the low data resource setting of AL. Existing approaches include semi-supervised learning (McCallum and Nigam, 1998; Tomanek and Hahn, 2009; Dasgupta and Ng, 2009; Yu et al., 2022), weak supervision (Ni et al., 2019; Qian et al., 2020; Brantley et al., 2020; Zhang et al., 2022a) and data augmentation (Zhang et al., 2020; Zhao et al., 2020; Hu and Neubig, 2021), with the most prevalent approach currently to be transfer learning from pretrained language models (Ein-Dor et al., 2020; Margatina et al., 2021; Tamkin et al., 2022). Recently, Margatina et al. (2022) showed large performance gains by adapting the pretrained language model to the task using the unlabeled data of the pool (i.e., task adaptive pretraining by Gururangan et al. (2020)). The authors also proposed an adaptive fine-tuning technique to account for the varying size of $\mathcal{D}_{\text{lab}}$ showing extra increase in $\mathcal{D}_{\text{test}}$ performance.

Still, there is room for improvement in this rather

[3] In AL research, a fully labeled dataset is typically *treated* as an *unlabeled* $\mathcal{D}_{\text{pool}}$ by entirely ignoring its labels, while in reality we *do* have access to them. Hence, the labels implicitly play a role in the design of the AL experiment. We analyze our criticism to this seemingly "random sampling" approach to form the seed dataset in §4.2.

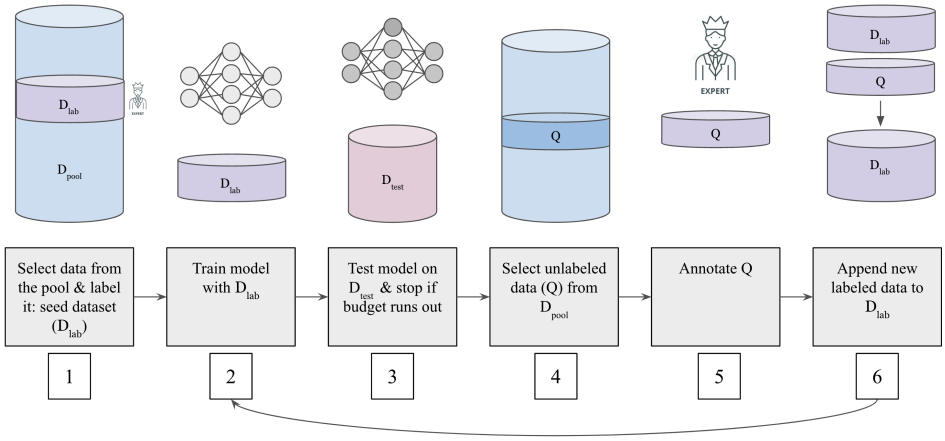


Figure 2: Distinct steps of the active learning loop (1–6). We use blue for the unlabeled data, purple for the labeled data and red for the (labeled) test data.

under-explored area. Especially now, state-of-the-art NLP pretrained language models consist of many millions or even billions of parameters. In AL we often deal with a small $\mathcal{D}_{\text{lab}}$ of a few hundred examples, thus adapting the training strategy is not trivial.

2.4 Data Acquisition

The data acquisition step (Figure 2: 4) is probably the core of the AL process and can be performed in various ways.[4]

Zhang et al. (2022d) provide a thorough literature review of query strategies, dividing them into two broad families. The first is based on *informativeness*, and methods in this family treat each candidate instance individually, assign a score and select the top (or bottom) instances based on the ranking of the scores. Major sub-categories of methods that belong in the informativeness family are uncertainty sampling (Lewis and Gale, 1994; Culotta and Mccallum, 2005; Zhang and Plank, 2021; Schröder et al., 2022), divergence-based algorithms (Ducoffe and Precioso, 2018; Margatina et al., 2021; Zhang et al., 2022b), disagreement-based (Seung et al., 1992; Houlsby et al., 2011; Gal et al., 2017; Siddhant and Lipton, 2018; Kirsch et al., 2019; Zeng and Zubiaga, 2023), gradient-based (Settles et al., 2007; Settles and Craven, 2008) and performance prediction (Roy and Mccallum, 2001; Konyushkova et al., 2017; Bachman et al., 2017; Liu et al., 2018).

The second family is representativeness and takes into account how instances of the pool correlate with each other, in order to avoid sampling bias harms from treating each instance individually. Density-based methods choose the most representative instances of the unlabeled pool (Ambati et al., 2010; Zhao et al., 2020; Zhu et al., 2008), while others opt for discriminative data points that differ from the already labeled dataset (Gissin and Shalev-Shwartz, 2019; Erdmann et al., 2019). A commonly adopted category in this family is batch diversity, where algorithms select a batch of diverse data points from the pool at each iteration (Brinker, 2003; Bodó et al., 2011; Zhu et al., 2008; Geifman and El-Yaniv, 2017; Zhdanov, 2019; Yu et al., 2022), with core-set (Sener and Savarese, 2018) to be the most common approach.

Naturally, there are hybrid acquisition functions that combine informativeness and representativeness (Yuan et al., 2020; Ash et al., 2020; Shi et al., 2021). Still, among the aforementioned methods there is not a universally superior acquisition function that consistently outperforms all others. Thus, which data to acquire is an active area of research.

2.5 Data Annotation

Once an acquisition function is applied to $\mathcal{D}_{\text{pool}}$, a subset Q is chosen, and the obtained unlabeled data is subsequently forwarded to human annotators for annotation (Figure 2: 5). In the context of simulation-based active learning, this aspect is not the primary focus since the labels for the actively acquired batch are *already* available. However, a

[4] In literature, the terms *data selection method*, *query strategy* and *acquisition function* are often used interchangeably.

question that naturally arises is: *Are all examples equally easy to annotate?* In simulation, all instances take equally long to label. This does not account for the fact that hard instances for the classifier are often hard for humans as well (Hachey et al., 2005; Baldridge and Osborne, 2004), therefore the current experimental setting is limiting and research for cost-aware selection strategies (Donmez and Carbonell, 2008; Tomanek and Hahn, 2010; Wei et al., 2019) is required. This would include explicit exploration of the synergies between random or actively acquired data and annotator expertise (Baldridge and Palmer, 2009).

2.6 Stopping Criterion

Finally, another active area of research is to develop effective methods for stopping AL (Figure 2: 3). In simulation, we typically decide as a budget a number of examples or a percentage of $\mathcal{D}_{\text{pool}}$ up to which we "aford" to annotate. However, in both research and real world applications, it is not clear if the model performance has reached a plateau. The stopping criterion should not be pre-defined by a heuristic, but rather a product of a well-designed experimental setting (Vlachos, 2008; Tomanek and Hahn, 2010; Ishibashi and Hino, 2020; Pullar-Strecker et al., 2022; Hacohen et al., 2022; Kurlandski and Bloodgood, 2022).[5]

3 The Fine Print

Previously, we presented specific challenges across different steps in the AL loop that researchers and practitioners need to address. Still, these challenges have long been attracting the attention of the research community. Interestingly, there are more caveats, that someone with no AL experience might have never encountered or even imagined. Hence, in this section we aim to unveil several such small details that still remain unexplored.

3.1 Hyperparameter Tuning

A possibly major issue of the current academic status quo in AL, is that researchers often do not tune the models-in-the-loop. This is mostly due to limitations related to time and compute constrains. For instance, a paper that proposes a new acquisition function would be required to run experiments for multiple baselines, iterations, random seeds and datasets. For example, a modest experiment including $a = 5$ acquisition functions, $i = 10$ AL iterations, $n = 5$ random seeds and $d = 5$ datasets, would reach an outstanding number of minimum $a \times i \times n \times d = 1,250$ trained models in total. This makes it rather hard to perform hyperparameter tuning of all these models in every AL loop, so it is the norm to use the same model architecture and hyperparameters to train all models.

In reality, practitioners that want to use AL, apply it *once*. Therefore, they most likely afford to tune the one and only model-in-the-loop. The question that arises then, is "*do the findings of AL experiments that do not tune the models generalize to scenarios where all models-in-the-loop are tuned*"? In other words, if an AL algorithm A performs better than B according to an experimental finding, would this be the case if we applied hyperparameter tuning to the models of both algorithms? Wouldn't it be possible that, with another configuration of hyperparameters, B performed better in the end?

3.2 Model Stability

In parallel, another undisclosed detail is what researchers do when the models-in-the-loop are unstable (i.e., *crash*). This essentially means that for some reason the optimisation of the model might fail and the model never converges leading to extremely poor predictive performance. Perhaps before the deep learning era such a problem did not exist, but now it is a likely phenomenon.

Dodge et al. (2020) showed that many fine-tuning experiments diverged part of the way through training especially on small datasets. AL is by definition connected with low-data resource settings, as the gains of data efficiency are meaningful in the scenario when labeled data is scarce. In light of this challenge, there is no consensus as to what an AL researcher or practitioner should do to alleviate this problem. One can choose to re-train the model with a different random seed, or do nothing. Though, it is non-trivial under which condition one should choose to re-train the model, since it is common that not always test performance improves from one AL iteration to the next.

Furthermore, there is currently no study that explores how much AL algorithms, that use the model-in-the-loop for acquisition, suffer by this problem. For instance, consider an uncertainty-based AL algorithm that uses the predictive proba-

[5] Unless of course the actual budget is spent, where in real world settings this is effectively the stopping criterion.

bility distribution of the model to select the most uncertain data points from the pool. If the model crashes, then its uncertainty estimates are not meaningful, thus the data acquisition function does not work as expected. In effect, the sampling method turns to a uniform distribution (i.e., the random sampling baseline).

3.3 Active Learning Evaluation

Another important challenge is the evaluation framework for AL. Evaluating the *actual* contribution of an AL method against its competitors would require to perform the same iterative *train-acquire-annotate* experiment (Figure 1) for all AL methods in the exact same data setting and with real human annotations. Certainly, such a laborious and expensive process is prohibitive for academic research, which is why we perform simulations by treating an *already* labeled and open-source dataset as a pool of unlabeled data.

Still, even if we were able to perform the experiments in real life, it is not trivial how to properly define when one method is better than another. This is because AL experiments include multiple rounds of annotation, thus multiple trained models and multiple scores in the test set(s). In cases with no clear difference between the algorithms compared, how should we do a fair comparison?

Previous work presents tables comparing the test set performance of the last model, often ignoring performance in previous loops (Prabhu et al., 2019; Mussmann et al., 2020). The vast majority of previous work though uses plots to visualize the performance over the AL iterations (Lowell et al., 2019; Ein-Dor et al., 2020) and in some cases offer a more detailed visualization with the variance due to the random seeds (Yuan et al., 2020; Kirsch et al., 2021; Margatina et al., 2021).

3.4 The Test of Time

Settles (2009) eloquently defines the "test of time" problem that AL faces: "*A training set built in cooperation with an active learner is inherently tied to the model that was used to generate it (i.e., the class of the model selecting the queries). Therefore, the labeled instances are a biased distribution, not drawn i.i.d. from the underlying natural density. If one were to change model classes—as we often do in machine learning when the state of the art advances—this training set may no longer be as useful to the new model class*".

Several years later, in the deep learning era, Lowell et al. (2019) indeed corroborates this concern. They demonstrate that a model from a certain family (e.g., convolution neural networks) might perform better when trained with a random subset of a pool, than an actively acquired dataset with a model of a different family (e.g., recurrent neural networks). Interestingly, Jelenić et al. (2023) recently showed that AL methods with similar acquisition sequences produce highly transferable datasets regardless of the model architecture. Related to the "test of time" challenge, it is rarely investigated whether the training data actively acquired with one model will confer benefits if used to train a second model (as compared to randomly sampled data from the same pool). Given that datasets often outlive learning algorithms, this is an important practical consideration (Baldridge and Osborne, 2004; Lowell et al., 2019; Shelmanov et al., 2021).

4 Active Learning in *Simulated* vs. Real World Settings

> *Is it truly logical to consider an already cleaned (preprocessed), typically published open-source labeled dataset as an unlabeled data pool for pool-based active learning simulation, with the expectation that any conclusions drawn will be applicable to real-world scenarios?*

The convenience and scalability of simulation make it an undoubtedly appealing approach for advancing machine learning research. In NLP, when tackling a specific task, for instance summarization, researchers often experiment with the limited availability of labeled summarization datasets, aiming to gain valuable insights and improve summarization models across various domains and languages. While this approach may not be ideal, it is a practical solution. *What makes the sub-field of active learning different?*

Admittedly, progress has, and will be made in AL research by leveraging simulation environments, similar to other areas within machine learning. Thus, there is no inherent requirement for a radically different approach in AL. We believe that simulating AL is indispensable for developing new methods and advancing the state-of-the-art.

Nonetheless, we argue that a slight distinction should be taken into account. AL is an iterative process that aims to obtain the smallest possible amount of labeled *data* given a substantially larger

pool of unlabeled data for maximizing predictive performance on a given task. The difference between developing models and constructing datasets lies in the fact that if a model is poorly trained, it can simply be retrained. Conversely, in AL, there exists a finite budget for acquiring annotations, and once it is expended, *there is no going back*. Consequently, we must have confidence that the AL state-of-the-art established through research simulations will perform equally well in practical applications.

Given these considerations, we advocate for a more critical approach to conducting simulation AL experiments. We should be addressing all the challenges (§2) and the experimental limitations (§3) discussed previously, while acknowledging the disparities between the simulation environment and real-world applications (§4.1). Given that datasets tend to outlast models (Lowell et al., 2019), we firmly believe that it is crucial to ensure the trustworthiness of AL research findings and their generalizability to real-world active data collection. This will contribute to the generation of high-quality datasets that stand the test of time (§3.4).

4.1 Simulation as a *Lower* Bound of Active Learning

The distribution gap between benchmark datasets in common ML tasks and data encountered in a real world production setting is well known (Bengio et al., 2020; Koh et al., 2021; Wang and Deng, 2018; Yin et al., 2021).

High Quality Data It is common practice for researchers to carefully curate the data to be labeled properly, often collecting multiple human annotations per example and discarding instances with disagreeing labels. When datasets are introduced in papers published in prestigious conferences or journals, it is expected that they should be of the highest quality, with an in-depth analysis of its data collection procedure, label distribution and other statistics. Nonetheless, it is important to acknowledge that such datasets may not encompass the entire spectrum of language variations encountered in real-world environments (Yin et al., 2021). Consequently, it remains uncertain whether an AL algorithm would generalize effectively to unfiltered raw data. Specifically, we hypothesize that the filtered data would be largely *more homogeneous* than the initial "pool". Assuming that the simulation $\mathcal{D}_{\text{pool}}$ is a somewhat homogeneous dataset, we can expect that *any* subset of data points drawn from it would, consequently, be more or less identical.[6] Therefore, if we train a model in each such subset, we would expect to obtain similar performance on test data due to the similarity between the training sets. From this perspective, random (uniform) sampling from a homogeneous pool can be considered a rudimentary form of diversity sampling.

Low Quality Data In contrast, it is possible that a publicly available dataset used for AL research may contain data of inferior quality, characterized by outliers such as repetitive instances, inadequate text filtering, incorrect labels, and implausible examples, among others. In such cases, an AL acquisition strategy, particularly one based on model uncertainty, may consistently select these instances for labeling due to their high level of data difficulty and uncertainty. Previous studies (Karamcheti et al., 2021; Snijders et al., 2023) have demonstrated the occurrence of this phenomenon, which poses a significant challenge as it undermines the potential value of AL. In a real-world AL scenario, it is plausible to have a dedicated team responsible for assessing the quality of acquired data and discarding instances of subpar quality. However, within the confines of a simulation, such data filtering is typically absent from the researcher's perspective, leading to potentially misleading experimental outcomes. Snijders et al. (2023) tried to address this issue in a multi-source setting for the task of natural language inference, and showed that while uncertainty-based strategies perform poorly due to the acquisition of collective outliers, when outliers are removed (from the pool), AL algorithms exhibited a noteworthy recovery and outperformed random baselines.

4.2 Simulation as an *Upper* Bound of Active Learning

However, one might argue for the exact opposite.

Favored Design Choices Previously, we mentioned that when selecting the seed dataset (§2.2) we typically randomly sample data from $\mathcal{D}_{\text{pool}}$, while keeping the label distribution of the true training set.[7] Hence, a balanced seed dataset is typically obtained, given that most classification datasets tend to exhibit a balanced label distribution. In

[6] Here we do not hint that all textual instances of a dataset are actually identical, but that they are more similar between them compared to the larger pool that they were created from.

[7] The "true training set" is the original one used as the pool ($\mathcal{D}_{\text{pool}}$) by removing the labels.

effect, the label distribution of $\mathcal{D}_{\text{pool}}$ would also be balanced, setting a strict constraint for AL simulation, as the actual label distribution of the unlabeled data should in reality be *unknown*. In other words, such subtle choices in the experimental design can introduce bias, making the simulated settings more trivial than more challenging real world AL settings where there is uncertainty as to the quality and the label distribution of data crawled online, that typically constitute the unlabeled pool.

Temporal Drift & Model Mismatch Datasets intended for research purposes are often constructed within a fixed timeframe, with minimal consideration for temporal concept drift issues (Röttger and Pierrehumbert, 2021; Lazaridou et al., 2021; Margatina et al., 2023b). However, it is important to recognize that this may not align with real-world applications, where the data distribution undergoes changes over time. The utilization of random and standard splits, commonly employed in AL research, can lead to overly optimistic performance estimates (Søgaard et al., 2021), which may not generalize to the challenges presented by real-world scenarios. Consequently, practitioners should consider this limitation when designing their active learning experiments. Lowell et al. (2019) also raises several practical obstacles neglected in AL research, such as that the acquired dataset may be disadvantageous for training subsequent models, and concludes that academic investigations of AL typically omit key real-world considerations that might overestimate its utility.

4.3 Main Takeaways

In summary, there exist compelling arguments that support both perspectives: simulation can serve as a lower bound by impeding the true advancement of AL methods, or it can implicitly favor AL experimental design, thus providing an upper bound for evaluation. The validity of these arguments likely varies across different cases. We can claim with certainty that this simulation setting, as described in this paper, is a far from perfect framework to evaluate AL algorithms among them and against random sampling. Nevertheless, we hypothesize that the lower bound argument (§4.1) might be more truthful. It is conceivable that AL data selection approaches may exhibit similar performance levels, either due to a lack of variation and diversity in the sampled pool of data or due to the presence of outliers that are not eliminated during the iterations. Hence, we contend that *simulation can be perceived as a lower bound for AL performance*, which helps explain why AL methods struggle to surpass the performance of random sampling. We undoubtedly believe that we can only obtain such answers by *exploring the AL simulation space in depth and by performing thorough analysis and extensive experiments to contrast the two theories.*

4.4 Active Learning in the LLMs Era

The field of active learning holds considerable importance in the current era of Large Language Models (LLMs). AL has recently been explore as a framework to identify the most useful demonstrations for in-context learning with LLMs (Zhang et al., 2022c; Diao et al., 2023; Margatina et al., 2023a). Additionally, AL is inherently intertwined with data-driven approaches that underpin recent advancements in artificial intelligence, such as reinforcement learning from human feedback (RLHF) (Christiano et al., 2023; OpenAI, 2022, 2023; Bai et al., 2022a). AL and RLHF represent two distinct approaches that tackle diverse aspects of the overarching problem of AI alignment (Askell et al., 2021). AL primarily focuses on optimizing the data acquisition process by selectively choosing informative instances for labeling, primarily within supervised or semi-supervised learning paradigms. On the other hand, RLHF aims to train reinforcement learning agents by utilizing human feedback as a means to surmount challenges associated with traditional reward signals. Despite their disparate methodologies, both AL and RLHF emphasize the criticality of incorporating human involvement to enhance the performance of machine learning and AI systems. Through active engagement of humans in the training process, AL and RLHF contribute to the development of AI systems that exhibit greater alignment with human values and demonstrate enhanced accountability (Bai et al., 2022a,b; Ganguli et al., 2022; Glaese et al., 2022; Sun et al., 2023; Kim et al., 2023). Consequently, the synergistic relationship between these two approaches warrants further exploration, as it holds the potential to leverage AL techniques in order to augment the data efficiency and robustness of RLHF methods.

5 Guidelines for Future Work

Given the inherent limitations of simulated AL settings, we propose guidelines to improve trustworthiness and robustness in AL research.

Transparency Our first recommendation is a call for transparency, which essentially means to *report everything* (Dodge et al., 2019). Every detail of the experimental setup, the implementation and the results, would be extremely helpful to properly evaluate the soundness of the experiments. We urge AL researchers to make use of the Appendix (or other means such as more detailed technical reports) to communicate interesting (or not) findings and problems, so that all details (§3) are accessible.

Thorough Experimental Settings We aim to incentivize researchers to thoughtfully consider ethical and practical aspects in their experimental settings. It is crucial to compare a wide range of algorithms, striving for generalizable results and findings across datasets, tasks, and domains. Moreover, we endorse research endeavors that aim to simulate more realistic settings for AL, such as exploration of AL across multiple domains (Longpre et al., 2022; Snijders et al., 2023). Additionally, we advocate for investigations into active learning techniques for languages beyond English, as the prevailing body of research predominantly focuses on English datasets (Bender, 2011).

Evaluation Protocol We strongly encourage researchers to prioritize the establishment of fair comparisons among different methods and to provide extensive presentation of results, including the consideration of variance across random seeds, in order to ensure robustness and reliability of findings. Generally, we argue that there is room for improvement of the active learning evaluation framework and we should explore approaches from other fields that promote more rigorous experimental and evaluation frameworks (Artetxe et al., 2020).

Analysis We place additional emphasis on the requirement of conducting comprehensive analysis of AL results. It is imperative to delve into the nuances of how different AL algorithms diverge and the extent of similarity (or dissimilarity) among the actively acquired datasets. It is incumbent upon AL research papers to extend beyond the results section and include an extensive analysis component, which provides deeper insights and understanding, as in Ein-Dor et al. (2020); Yuan et al. (2020); Margatina et al. (2021); Zhou et al. (2021); Snijders et al. (2023), among others. If we aim to unveil why an AL algorithm fails to outperform another (or the random baseline), we need to understand which data it selected in the first place, and why.

Reproducibility Reproducing AL experiments can be challenging due to the complex nature of a typical AL experiment, involving multiple rounds of model training and evaluation, which can be computationally demanding. However, we strongly advocate for practitioners and researchers to prioritize the release of their code and provide comprehensive instructions for future researchers aiming to build upon their work. By making code and associated resources available, the research community can foster transparency, facilitate replication, and enable further advancements in AL methodologies.

Efficiency Finally, we propose the release of actively acquired datasets generated by different AL algorithms, which would greatly contribute to data-centric research and interpretability aspects of AL. Particularly when employing AL with large-scale models, it becomes crucial to establish the actively acquired data from other studies as baselines, rather than re-running the entire process from the beginning. Such an approach would not only enhance transparency, but also promote efficiency and eco-friendly practices within the research community.

6 Conclusion

In this position paper, we examine the numerous challenges encountered throughout the various stages of the active learning pipeline. Additionally, we provide a comprehensive overview of the often-overlooked limitations within the AL research community, with the intention of illuminating obscure experimental design choices. Furthermore, we delve into a thorough exploration of the limitations associated with simulation in AL, engaging in a critical discussion regarding its potential as either a lower or upper bound on AL performance. Lastly, we put forth guidelines for future research directions, aimed at enhancing the robustness and credibility of AL research for effective real-world applications. This perspective is particularly timely, particularly considering the notable advancements in modeling within the field NLP (e.g., ChatGPT[8], Claude[9]) . These advancements have resulted in a shift of emphasis towards a more data-centric approach in machine learning research, emphasizing the significance of carefully selecting relevant data to enhance models and ensure their alignment with human values.

[8] https://openai.com/blog/chatgpt

[9] https://www.anthropic.com/index/introducing-claude

Limitations

In this position paper, we have strived to provide a comprehensive overview, acknowledging that there may be relevant research papers that have inadvertently escaped our attention. While we have made efforts to include a diverse range of related work from various fields, such as machine learning and computer vision, it is important to note that our analysis predominantly focuses on AL papers presented at NLP conferences. Moreover, it is worth mentioning that the majority, if not all, of the AL papers examined and referenced in this survey are centered around the English language, thereby limiting the generalizability and applicability of our findings and critiques to other languages and contexts. We wish to emphasize that the speculations put forth in this position paper carry no substantial risks, as they are substantiated by peer-reviewed papers, and our hypotheses (§4) are explicitly stated as such, representing conjectures rather than definitive findings regarding the role of simulation in AL research. We sincerely hope that this paper stimulates robust discussions and undergoes thorough scrutiny by experts in the field, with the ultimate objective of serving as a valuable guideline for AL researchers, particularly graduate students, seeking to engage in active learning research. Above all, *we earnestly urge researchers equipped with the necessary resources to conduct experiments and analyses that evaluate our hypotheses, striving to bridge the gap between research and real-world settings in the context of active learning.*

Acknowledgements

We would like to thank the anonymous reviewers for their insightful feedback. Both authors are supported by an Amazon Alexa Fellowship.

References

Vamshi Ambati, Stephan Vogel, and Jaime Carbonell. 2010. Active semi-supervised learning for improving word alignment. In *Proceedings of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing*, ALNLP '10, pages 10–17, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, Dani Yogatama, Gorka Labaka, and Eneko Agirre. 2020. A call for more rigor in unsupervised cross-lingual learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7375–7388, Online. Association for Computational Linguistics.

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2020. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. A general language assistant as a laboratory for alignment.

Philip Bachman, Alessandro Sordoni, and Adam Trischler. 2017. Learning algorithms for active learning. In *Proceedings of the International Conference on Machine Learning*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022b. Constitutional ai: Harmlessness from ai feedback.

Jason Baldridge and Miles Osborne. 2004. Active learning and the total cost of annotation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 9–16, Barcelona, Spain. Association for Computational Linguistics.

Jason Baldridge and Alexis Palmer. 2009. How well does active learning *actually* work? Time-based evaluation of cost-reduction strategies for language documentation. In *Proceedings of the 2009 Conference on*

Empirical Methods in Natural Language Processing, pages 296–305, Singapore. Association for Computational Linguistics.

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the AI: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678.

Max Bartolo, Tristan Thrush, Sebastian Riedel, Pontus Stenetorp, Robin Jia, and Douwe Kiela. 2022. Models in the loop: Aiding crowdworkers with generative annotation assistants. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3754–3767, Seattle, United States. Association for Computational Linguistics.

Emily M. Bender. 2011. On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6.

Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Nan Rosemary Ke, Sebastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. 2020. A meta-transfer objective for learning to disentangle causal mechanisms. In *International Conference on Learning Representations*.

Zalán Bodó, Zsolt Minier, and Lehel Csató. 2011. Active learning with clustering. In *Proceedings of the Active Learning and Experimental Design workshop In conjunction with AISTATS 2010*, volume 16, pages 127–139.

Kianté Brantley, Amr Sharaf, and Hal Daumé III. 2020. Active imitation learning with noisy guidance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2093–2105, Online. Association for Computational Linguistics.

Klaus Brinker. 2003. Incorporating diversity in active learning with support vector machines. In *Proceedings of the International Conference on Machine Learning*, pages 59–66.

Aditi Chaudhary, Antonios Anastasopoulos, Zaid Sheikh, and Graham Neubig. 2021. Reducing confusion in active learning for part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 9:1–16.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2023. Deep reinforcement learning from human preferences.

David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. 1996. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4(1):129–145.

Aron Culotta and Andrew Mccallum. 2005. Reducing labeling effort for structured prediction tasks. In *Association for the Advancement of Artificial Intelligence*.

Aswarth Abhilash Dara, Josef van Genabith, Qun Liu, John Judge, and Antonio Toral. 2014. Active learning for Post-Editing based incrementally retrained MT. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 185–189, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sajib Dasgupta and Vincent Ng. 2009. Mine the easy, classify the hard: A semi-supervised approach to automatic sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 701–709, Suntec, Singapore. Association for Computational Linguistics.

Shizhe Diao, Pengcheng Wang, Yong Lin, and Tong Zhang. 2023. Active prompting with chain-of-thought for large language models.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194, Hong Kong, China. Association for Computational Linguistics.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping.

Pinar Donmez and Jaime G. Carbonell. 2008. Proactive learning: Cost-sensitive active learning with multiple imperfect oracles. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, page 619–628, New York, NY, USA. Association for Computing Machinery.

Melanie Ducoffe and Frédéric Precioso. 2018. Adversarial active learning for deep networks: a margin based approach. *CoRR*, abs/1802.09841.

Liat Ein-Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2020. Active Learning for BERT: An Empirical Study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7949–7962, Online. Association for Computational Linguistics.

Alexander Erdmann, David Joseph Wrisley, Benjamin Allen, Christopher Brown, Sophie Cohen-Bodénès, Micha Elsner, Yukun Feng, Brian Joseph, Béatrice

Joyeux-Prunel, and Marie-Catherine de Marneffe. 2019. Practical, efficient, and customizable active learning for named entity recognition in the digital humanities. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2223–2234.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep Bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1183–1192. PMLR.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned.

Yonatan Geifman and Ran El-Yaniv. 2017. Deep active learning over the long tail. *CoRR*, abs/1711.00941.

Daniel Gissin and Shai Shalev-Shwartz. 2019. Discriminative active learning.

Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. 2022. Improving alignment of dialogue agents via targeted human judgements.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Ben Hachey, Beatrice Alex, and Markus Becker. 2005. Investigating the effects of selective sampling on the annotation task. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 144–151, Ann Arbor, Michigan. Association for Computational Linguistics.

Guy Hacohen, Avihu Dekel, and Daphna Weinshall. 2022. Active learning on a budget: Opposite strategies suit high and low budgets. *CoRR*, abs/2202.02794.

Gholamreza Haffari, Maxim Roy, and Anoop Sarkar. 2009. Active learning for statistical phrase-based machine translation. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 415–423.

Andrea Horbach and Alexis Palmer. 2016. Investigating active learning for short-answer scoring. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 301–311, San Diego, CA. Association for Computational Linguistics.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian active learning for classification and preference learning. *ArXiv*.

Junjie Hu and Graham Neubig. 2021. Phrase-level active learning for neural machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1087–1099, Online. Association for Computational Linguistics.

Rong Hu, Brian Mac Namee, and Sarah Jane Delany. 2010. Off to a good start: Using clustering to select the initial training set in active learning.

Hideaki Ishibashi and Hideitsu Hino. 2020. Stopping criterion for active learning based on deterministic generalization bounds. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 386–397. PMLR.

Fran Jelenić, Josip Jukić, Nina Drobac, and Jan Šnajder. 2023. On dataset transferability in active learning for transformers.

Jaeho Kang, Kwang Ryel Ryu, and Hyuk chul Kwon. 2004. Using cluster-based sampling to select initial training set for active learning in text classification. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*.

Siddharth Karamcheti, Ranjay Krishna, Li Fei-Fei, and Christopher Manning. 2021. Mind your outliers! investigating the negative impact of outliers on active learning for visual question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7265–7281, Online. Association for Computational Linguistics.

Twin Karmakharm, Nikolaos Aletras, and Kalina Bontcheva. 2019. Journalist-in-the-loop: Continuous learning as a service for rumour analysis. In *Proceedings of the 2019 Conference on Empirical Methods*

in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations, pages 115–120, Hong Kong, China. Association for Computational Linguistics.

Jungo Kasai, Kun Qian, Sairam Gurajada, Yunyao Li, and Lucian Popa. 2019. Low-resource deep entity resolution with transfer and active learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5851–5861, Florence, Italy. Association for Computational Linguistics.

Nataliia Kees, Michael Fromm, Evgeniy Faerman, and Thomas Seidl. 2021. Active learning for argument strength estimation. In *Proceedings of the Second Workshop on Insights from Negative Results in NLP*, pages 144–150, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Sungdong Kim, Sanghwan Bae, Jamin Shin, Soyoung Kang, Donghyun Kwak, Kang Min Yoo, and Minjoon Seo. 2023. Aligning large language models through synthetic feedback.

Andreas Kirsch, Tom Rainforth, and Yarin Gal. 2021. Test distribution-aware active learning: A principled approach against distribution shift and outliers.

Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. 2019. BatchBALD: Efficient and diverse batch acquisition for deep bayesian active learning. In *Neural Information Processing Systems*, pages 7026–7037.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5637–5664. PMLR.

Ksenia Konyushkova, Raphael Sznitman, and Pascal Fua. 2017. Learning active learning from data. In I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett, editors, *Advances in Neural Information Processing Systems*.

Luke Kurlandski and Michael Bloodgood. 2022. Impact of stop sets on stopping active learning for text classification. *CoRR*, abs/2201.05460.

Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomáš Kočiský, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. 2021. Mind the gap: Assessing temporal generalization in neural language models. In *Advances in Neural Information Processing Systems*.

David D. Lewis and William A. Gale. 1994. A sequential algorithm for training text classifiers. In *In Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Ming Liu, Wray Buntine, and Gholamreza Haffari. 2018. Learning how to actively learn: A deep imitation learning approach. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1874–1883, Melbourne, Australia. Association for Computational Linguistics.

Shayne Longpre, Julia Reisler, Edward Greg Huang, Yi Lu, Andrew Frank, Nikhil Ramesh, and Chris DuBois. 2022. Active learning over multiple domains in natural language tasks.

David Lowell, Zachary C. Lipton, and Byron C. Wallace. 2019. Practical obstacles to deploying active learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 21–30, Hong Kong, China. Association for Computational Linguistics.

Katerina Margatina, Loic Barrault, and Nikolaos Aletras. 2022. On the importance of effectively adapting pretrained language models for active learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 825–836, Dublin, Ireland. Association for Computational Linguistics.

Katerina Margatina, Timo Schick, Nikolaos Aletras, and Jane Dwivedi-Yu. 2023a. Active learning principles for in-context learning with large language models.

Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. 2021. Active learning by acquiring contrastive examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 650–663, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Katerina Margatina, Shuai Wang, Yogarshi Vyas, Neha Anna John, Yassine Benajiba, and Miguel Ballesteros. 2023b. Dynamic benchmarking of masked language models on temporal concept drift with multiple views.

In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2881–2898, Dubrovnik, Croatia. Association for Computational Linguistics.

Andrew McCallum and Kamal Nigam. 1998. Employing em and pool-based active learning for text classification. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, page 350–358, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Akiva Miura, Graham Neubig, Michael Paul, and Satoshi Nakamura. 2016. Selecting syntactic, non-redundant segments in active learning for machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 20–29, Stroudsburg, PA, USA. Association for Computational Linguistics.

Stephen Mussmann, Robin Jia, and Percy Liang. 2020. On the importance of adaptive data collection for extremely imbalanced pairwise tasks.

Ansong Ni, Pengcheng Yin, and Graham Neubig. 2019. Merging weak and active supervision for semantic parsing. In *AAAI Conference on Artificial Intelligence*.

OpenAI. 2022. Chatgpt.

OpenAI. 2023. Gpt-4 technical report.

Ameya Prabhu, Charles Dognin, and Maneesh Singh. 2019. Sampling bias in deep active classification: An empirical study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, pages 4056–4066.

Zac Pullar-Strecker, Katharina Dost, Eibe Frank, and Jörg Wicker. 2022. Hitting the target: stopping active learning at the cost-based optimum. *Machine Learning*, pages 1–19.

Kun Qian, Poornima Chozhiyath Raman, Yunyao Li, and Lucian Popa. 2020. Learning structured representations of entity names using Active Learning and weak supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6376–6383, Online. Association for Computational Linguistics.

Kun Qian, Lucian Popa, and Prithviraj Sen. 2017. Active learning for Large-Scale entity resolution. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1379–1388. ACM.

Paul Röttger and Janet Pierrehumbert. 2021. Temporal adaptation of BERT and performance on downstream document classification: Insights from social media. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2400–2412, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nicholas Roy and Andrew Mccallum. 2001. Toward optimal active learning through monte carlo estimation of error reduction. In *in Proceedings of the International Conference on Machine Learning*.

Christopher Schröder, Lydia Müller, Andreas Niekler, and Martin Potthast. 2023. Small-text: Active learning for text classification in python. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 84–95, Dubrovnik, Croatia. Association for Computational Linguistics.

Christopher Schröder and Andreas Niekler. 2020. A survey of active learning for text classification using deep neural networks. *CoRR*, abs/2008.07267.

Christopher Schröder, Andreas Niekler, and Martin Potthast. 2022. Revisiting uncertainty-based query strategies for active learning with transformers. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2194–2203, Dublin, Ireland. Association for Computational Linguistics.

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*.

Burr Settles. 2009. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Burr Settles and Mark Craven. 2008. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1070–1079, Honolulu, Hawaii. Association for Computational Linguistics.

Burr Settles, Mark Craven, and Soumya Ray. 2007. Multiple-instance active learning. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc.

H. S. Seung, M. Opper, and H. Sompolinsky. 1992. Query by committee. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, page 287–294, New York, NY, USA. Association for Computing Machinery.

Artem Shelmanov, Dmitri Puzyrev, Lyubov Kupriyanova, Denis Belyakov, Daniil Larionov, Nikita Khromov, Olga Kozlova, Ekaterina Artemova, Dmitry V. Dylov, and Alexander Panchenko. 2021. Active learning for sequence tagging with deep pre-trained models and Bayesian uncertainty estimates. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1698–1712, Online. Association for Computational Linguistics.

Yanyao Shen, Hyokun Yun, Zachary Lipton, Yakov Kronrod, and Animashree Anandkumar. 2017. Deep

active learning for named entity recognition. In *Proceedings of the Workshop on Representation Learning for NLP*, pages 252–256.

Tianze Shi, Adrian Benton, Igor Malioutov, and Ozan İrsoy. 2021. Diversity-aware batch active learning for dependency parsing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2616–2626, Online. Association for Computational Linguistics.

Aditya Siddhant and Zachary C. Lipton. 2018. Deep Bayesian active learning for natural language processing: Results of a large-scale empirical study. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2904–2909, Brussels, Belgium. Association for Computational Linguistics.

Ard Snijders, Douwe Kiela, and Katerina Margatina. 2023. Investigating multi-source active learning for natural language inference. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2187–2209, Dubrovnik, Croatia. Association for Computational Linguistics.

Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, and Katja Filippova. 2021. We need to talk about random splits. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1823–1832, Online. Association for Computational Linguistics.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2023. Principle-driven self-alignment of language models from scratch with minimal human supervision.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Alex Tamkin, Dat Pham Nguyen, Salil Deshpande, Jesse Mu, and Noah Goodman. 2022. Active learning helps pretrained models learn the intended task. In *Advances in Neural Information Processing Systems*.

Katrin Tomanek and Udo Hahn. 2009. Semi-supervised active learning for sequence labeling. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1039–1047, Suntec, Singapore. Association for Computational Linguistics.

Katrin Tomanek and Udo Hahn. 2010. A comparison of models for cost-sensitive active learning. In *Coling 2010: Posters*, pages 1247–1255, Beijing, China. Coling 2010 Organizing Committee.

Katrin Tomanek, Florian Laws, Udo Hahn, and Hinrich Schütze. 2009. On proper unit selection in active learning: Co-selection effects for named entity recognition. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 9–17, Boulder, Colorado. Association for Computational Linguistics.

Andreas Vlachos. 2006. Active annotation. In *Proceedings of the Workshop on Adaptive Text Extraction and Mining (ATEM 2006)*.

Andreas Vlachos. 2008. A stopping criterion for active learning. *Computer Speech & Language*, 22(3):295–312.

Eric Wallace, Adina Williams, Robin Jia, and Douwe Kiela. 2022. Analyzing dynamic adversarial training data in the limit. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 202–217, Dublin, Ireland. Association for Computational Linguistics.

Mei Wang and Weihong Deng. 2018. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153.

Qiang Wei, Yukun Chen, Mandana Salimi, Joshua C Denny, Qiaozhu Mei, Thomas A Lasko, Qingxia Chen, Stephen Wu, Amy Franklin, Trevor Cohen, and Hua Xu. 2019. Cost-aware active learning for named entity recognition in clinical text. *Journal of the American Medical Informatics Association*, 26(11):1314–1322.

Wenpeng Yin, Shelby Heinecke, Jia Li, Nitish Shirish Keskar, Michael Jones, Shouzhong Shi, Stanislav Georgiev, Kurt Milich, Joseph Esposito, and Caiming Xiong. 2021. Combining data-driven supervision with human-in-the-loop feedback for entity resolution.

Yue Yu, Lingkai Kong, Jieyu Zhang, Rongzhi Zhang, and Chao Zhang. 2022. AcTune: Uncertainty-based active self-training for active fine-tuning of pretrained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1422–1436, Seattle, United States. Association for Computational Linguistics.

Michelle Yuan, Hsuan-Tien Lin, and Jordan Boyd-Graber. 2020. Cold-start active learning through self-supervised language modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7935–7948, Online. Association for Computational Linguistics.

Michelle Yuan, Patrick Xia, Chandler May, Benjamin Van Durme, and Jordan Boyd-Graber. 2022. Adapting coreference resolution models through active

learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7533–7549, Dublin, Ireland. Association for Computational Linguistics.

Xia Zeng and Arkaitz Zubiaga. 2023. Active PETs: Active data annotation prioritisation for few-shot claim verification with pattern exploiting training. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 190–204, Dubrovnik, Croatia. Association for Computational Linguistics.

Mike Zhang and Barbara Plank. 2021. Cartography active learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 395–406, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rongzhi Zhang, Yue Yu, Pranav Shetty, Le Song, and Chao Zhang. 2022a. Prompt-based rule discovery and boosting for interactive weakly-supervised learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 745–758, Dublin, Ireland. Association for Computational Linguistics.

Rongzhi Zhang, Yue Yu, and Chao Zhang. 2020. SeqMix: Augmenting active sequence labeling via sequence mixup. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8566–8579, Online. Association for Computational Linguistics.

Shujian Zhang, Chengyue Gong, Xingchao Liu, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. 2022b. ALLSH: Active learning guided by local sensitivity and hardness. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1328–1342, Seattle, United States. Association for Computational Linguistics.

Yiming Zhang, Shi Feng, and Chenhao Tan. 2022c. Active example selection for in-context learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9134–9148, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhisong Zhang, Emma Strubell, and Eduard Hovy. 2022d. A survey of active learning for natural language processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Yuekai Zhao, Haoran Zhang, Shuchang Zhou, and Zhihua Zhang. 2020. Active learning approaches to enhancing neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1796–1806, Online. Association for Computational Linguistics.

Fedor Zhdanov. 2019. Diverse mini-batch active learning.

Yilun Zhou, Adithya Renduchintala, Xian Li, Sida Wang, Yashar Mehdad, and Asish Ghoshal. 2021. Towards understanding the behaviors of optimal deep active learning algorithms. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1486–1494. PMLR.

Jingbo Zhu, Huizhen Wang, Tianshun Yao, and Benjamin K Tsou. 2008. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 1137–1144, Manchester, UK. Coling 2008 Organizing Committee.

5.3 Impact

According to Google Scholar, the paper has received 7 citations as of May 2024. The propositions in our paper have influenced follow-up work (Ghose and Nguyen, 2024; Rouzegar and Makrehchi, 2024).

5.4 Discussion

The paper examines the challenges and limitations surrounding active learning (AL) in natural language processing (NLP) research. It surveys recent literature on AL, highlighting the growing interest in AL within the NLP community as language model pretraining advances. Despite its potential, AL algorithms are shown to fail to outperform random sampling, prompting questions about their fundamental assumptions. The paper identifies challenges within the AL loop, including problem definition, model training, data acquisition, and evaluation frameworks, emphasizing the need for a rigorous approach to conducting AL experiments and considering practical implications in real-world scenarios.

A significant portion of the paper focuses on the role of simulation in evaluating AL algorithms, framing it as both a lower and upper bound. Simulation is seen as a lower bound due to its limitations in replicating real-world scenarios accurately, potentially leading to misleading outcomes. However, it can also serve as an upper bound by imposing strict constraints on experimental settings, biasing results towards more straightforward scenarios. The main takeaway suggests that while simulation offers insights into AL performance, it may not fully capture the complexities of real-world scenarios, hindering the true advancement of AL methods.

In conclusion, the paper advocates for transparency, thorough experimental settings, and rigorous evaluation protocols in AL research to address the identified challenges and limitations. By providing guidelines for future research directions, including the exploration of AL techniques across various domains and languages, the paper aims to enhance the credibility and applicability of AL research in real-world settings. Overall, it highlights the importance of addressing the complexities of AL and considering practical implications to drive meaningful progress in the field, particularly within the context of advancing NLP research.

Chapter 6

Conclusion

This thesis introduces two approaches to improving active learning with (large) language models for text classification tasks, one study on using active learning algorithms to improve in-context learning with large language models, as well as a theory on what are the painpoints of active learning simulation and how to address them in the future. In Chapter 2, we introduce two techniques for fine-tuning pre-trained language models in a low-data resource setting of active learning Chapter and show substantial improvements with their application. Chapter 3 introduces CAL, a novel acquisition function for active learning that selects examples that are close to the model's decision boundary using the feature space of the hidden representations of the model to define which data points are most useful to be added to the training set. The focus of Chapter 4 is to explore how active learning algorithms can be applied to select demonstrations for in-context learning with large language models. Lastly, in Chapter 5, we criticize common practices in active learning simulation experiments and propose guidelines for future work on the field.

All presented algorithms are designed to be relatively simple, allowing for easy adaptation to various tasks and domains beyond NLP, even though the focus here is on text classification. While the models developed in this thesis are limited to pretrained language models like BERT (except for the study in Chapter 4), extending these models to newer large language models is relatively straightforward and has already been partially achieved by subsequent work. The theoretical part of the thesis, which discusses potential limitations in the experimental setup of active learning simulations, provides a thorough review of AL practices and proposes future improvements that are general enough to remain relevant amidst ongoing advancements in AL and NLP.

The key insights obtained from the exploration of active learning (AL) algorithms for

data-efficient language models are as follows:

- Model Adaptation in Increasing Dataset Size: When executing standard pool-based active learning with supervised learning methods, adapting the model to an increasing dataset size presents significant challenges. A static training protocol proves suboptimal as it must accommodate the evolving dataset size.
- Domain-Specific Adaptation: Adapting a pretrained language model to the specific domain of the task is unequivocally advantageous. This approach is particularly well-suited to the pool-based AL framework, where a substantial amount of unlabeled data is available beforehand, facilitating effective model domain adaptation.
- Criticality of Data Selection: The data selection phase within the AL pipeline is pivotal to the framework's overall success. Despite the absence of a universally superior method applicable across all tasks, domains, languages, and settings, prioritizing high-uncertainty and challenging examples near the model's decision boundary emerges as a highly promising strategy.
- Quality of Annotations: Ensuring the accuracy and high quality of annotations is paramount for AL performance. The presence of mislabeled data within the (small) training set can severely detract from performance, underscoring the necessity of allocating resources to secure high-quality annotations.
- Contrasts Between AL and In-Context Learning: The principles guiding multiple rounds of standard pool-based AL with a supervised learning model differ markedly from those of a single data selection step for demonstrations in in-context learning. In the former, high-uncertainty data proves beneficial, whereas in the latter, the critical factor is the semantic similarity of demonstrations to the test data, ensuring a semantically cohesive prompt.
- Discrepancies Between Simulations and Real-World Applications: Conducting AL simulations in a research context diverges substantially from real-world AL applications. It remains uncertain whether research-derived conclusions will generalize seamlessly to practical settings. We provide extensive arguments indicating that simulation results could represent either a lower or upper bound of real-world performance. Nevertheless, it is evident that several specific issues within simulation experiments need to be addressed, with experimental decisions requiring transparency and careful consideration.

These insights highlight the intricate nature of advancing active learning methodologies within the realm of natural language processing, raising several compelling avenues for

future investigation:

- How can contemporary large language models (LLMs) be fully leveraged within an active learning framework? Can they serve as models-in-the-loop without necessitating weight adjustments (i.e., no training)? Alternatively, can we exploit smaller, faster, and more cost-effective versions of these models?
- Is it feasible to employ LLMs as annotators to enhance simulation quality or mitigate the need for human intervention? What are the implications for reliability and the generation of unbiased outcomes?
- While we have addressed the active learning problem in the context of in-context learning, numerous unexplored dimensions remain. How can we devise efficient strategies for multiple-round active learning to select demonstrations? How can the selection of subsequent demonstrations be influenced by those previously chosen? Furthermore, how can these strategies be implemented in scenarios where access to the test set is restricted, thereby hindering the assurance of semantic similarity with the test data?

Active learning, although heavily reliant on model training and algorithmic modeling, is fundamentally a data-driven research domain. The ultimate goal extends beyond merely developing a well-performing, data-efficient model; it encompasses the creation of an actively acquired labeled dataset that endures the test of time. Given the enduring importance of labeled data, especially in light of the training requirements for contemporary large language models (LLMs), the role of active learning in generating high-quality datasets assumes heightened significance. As discussions increasingly center on responsible AI, regulatory frameworks, autonomous agents, and the imperative of aligning LLMs with human values, active learning emerges as a pivotal method for addressing these evolving challenges. Consequently, the advancement of active learning methodologies is imperative to navigate the complexities that lie ahead.

Bibliography

Ambati, V., Vogel, S., and Carbonell, J. (2010). Active semi-supervised learning for improving word alignment. In *Proceedings of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing*, ALNLP '10, pages 10–17, Stroudsburg, PA, USA. Association for Computational Linguistics.

Azeemi, A. H., Qazi, I. A., and Raza, A. A. (2023a). Representative subset selection for efficient fine-tuning in self-supervised speech recognition.

Azeemi, A. H., Qazi, I. A., and Raza, A. A. (2023b). Towards representative subset selection for self-supervised speech recognition.

Bachman, P., Sordoni, A., and Trischler, A. (2017). Learning algorithms for active learning. In *Proceedings of the International Conference on Machine Learning*.

Baldridge, J. and Osborne, M. (2004). Active learning and the total cost of annotation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 9–16, Barcelona, Spain. Association for Computational Linguistics.

Baldridge, J. and Palmer, A. (2009). How well does active learning *actually* work? Time-based evaluation of cost-reduction strategies for language documentation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 296–305, Singapore. Association for Computational Linguistics.

Bansal, P. and Sharma, A. (2023). Large language models as annotators: Enhancing generalization of nlp models at minimal cost. *arXiv preprint arXiv:2306.15766*.

Black, S., Biderman, S., Hallahan, E., Anthony, Q., Gao, L., Golding, L., He, H., Leahy, C., McDonell, K., Phang, J., Pieler, M., Prashanth, U. S., Purohit, S., Reynolds, L., Tow, J., Wang, B., and Weinbach, S. (2022). GPT-NeoX-20B: An open-source autoregressive language model. In Fan, A., Ilic, S., Wolf, T., and Gallé, M., editors, *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136, virtual+Dublin. Association for Computational Linguistics.

Bodó, Z., Minier, Z., and Csató, L. (2011). Active learning with clustering. In *Proceedings of the Active Learning and Experimental Design workshop In conjunction with AISTATS 2010*, volume 16, pages 127–139.

Brantley, K., Sharaf, A., and Daumé III, H. (2020). Active imitation learning with noisy guidance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2093–2105, Online. Association for Computational Linguistics.

Brinker, K. (2003). Incorporating diversity in active learning with support vector machines. In *Proceedings of the International Conference on Machine Learning*, pages 59–66.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Chaudhary, A., Anastasopoulos, A., Sheikh, Z., and Neubig, G. (2021). Reducing confusion in active learning for part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 9:1–16.

Chen, C., Wang, Y., Liao, L., Chen, Y., and Du, X. (2023). Real: A representative error-driven approach for active learning.

Chen, J., Lin, H., Han, X., Lu, Y., Jiang, S., Dong, B., and Sun, L. (2024). Few-shot named entity recognition via superposition concept discrimination. *arXiv preprint arXiv:2403.16463*.

Cohn, D. A., Ghahramani, Z., and Jordan, M. I. (1996). Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4(1):129–145.

Culotta, A. and Mccallum, A. (2005). Reducing labeling effort for structured prediction tasks. In *Association for the Advancement of Artificial Intelligence*.

Dara, A. A., van Genabith, J., Liu, Q., Judge, J., and Toral, A. (2014). Active learning for Post-Editing based incrementally retrained MT. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 185–189, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dasgupta, S. and Ng, V. (2009). Mine the easy, classify the hard: A semi-supervised approach to automatic sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 701–709, Suntec, Singapore. Association for Computational Linguistics.

Delaflor, M., Gendron, C., and Delgado-Solórzano, C. T. (2022). Reactin: The role of human feedback in reason-act prompting strategies with language models.

Deng, X., Wang, W., Feng, F., Zhang, H., He, X., and Liao, Y. (2023). Counterfactual active learning for out-of-distribution generalization. In *Annual Meeting of the Association for Computational Linguistics*, pages 11362–11377.

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 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Donmez, P. and Carbonell, J. G. (2008). Proactive learning: Cost-sensitive active

learning with multiple imperfect oracles. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, page 619–628, New York, NY, USA. Association for Computing Machinery.

Ducoffe, M. and Precioso, F. (2018). Adversarial active learning for deep networks: a margin based approach. *CoRR*, abs/1802.09841.

Ein-Dor, L., Halfon, A., Gera, A., Shnarch, E., Dankin, L., Choshen, L., Danilevsky, M., Aharonov, R., Katz, Y., and Slonim, N. (2020). Active Learning for BERT: An Empirical Study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7949–7962, Online. Association for Computational Linguistics.

Erdmann, A., Wrisley, D. J., Allen, B., Brown, C., Cohen-Bodénès, S., Elsner, M., Feng, Y., Joseph, B., Joyeux-Prunel, B., and de Marneffe, M.-C. (2019). Practical, efficient, and customizable active learning for named entity recognition in the digital humanities. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2223–2234.

Gal, Y., Islam, R., and Ghahramani, Z. (2017). Deep Bayesian active learning with image data. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1183–1192. PMLR.

Gao, L., Chaudhary, A., Srinivasan, K., Hashimoto, K., Raman, K., and Bendersky, M. (2023). Ambiguity-aware in-context learning with large language models. *arXiv preprint arXiv:2309.07900*.

Garg, I. and Roy, K. (2023). Samples with low loss curvature improve data efficiency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20290–20300.

Geifman, Y. and El-Yaniv, R. (2017). Deep active learning over the long tail. *CoRR*, abs/1711.00941.

Ghose, A. and Nguyen, E. (2024). On the fragility of active learners. *arXiv preprint arXiv:2403.15744*.

Gissin, D. and Shalev-Shwartz, S. (2019). Discriminative active learning.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Hachey, B., Alex, B., and Becker, M. (2005). Investigating the effects of selective sampling on the annotation task. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 144–151, Ann Arbor, Michigan. Association for Computational Linguistics.

Haffari, G., Roy, M., and Sarkar, A. (2009). Active learning for statistical phrase-based machine translation. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 415–423.

Hassan, S. and Alikhani, M. (2023). D-CALM: A dynamic clustering-based active learning approach for mitigating bias. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5540–5553, Toronto, Canada. Association for Computational Linguistics.

Houlsby, N., Huszár, F., Ghahramani, Z., and Lengyel, M. (2011). Bayesian active learning for classification and preference learning. *ArXiv*.

Hu, J. and Neubig, G. (2021). Phrase-level active learning for neural machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1087–1099, Online. Association for Computational Linguistics.

Hu, M., Zhang, Z., Zhao, S., Huang, M., and Wu, B. (2023a). Uncertainty in natural language processing: Sources, quantification, and applications.

Hu, Q., Guo, Y., Xie, X., Cordy, M., Ma, L., Papadakis, M., and Traon, Y. L. (2023b). Active code learning: Benchmarking sample-efficient training of code models.

Karisani, P., Karisani, N., and Xiong, L. (2022). Multi-view active learning for short text classification in user-generated data. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6441–6453.

Kasai, J., Qian, K., Gurajada, S., Li, Y., and Popa, L. (2019). Low-resource deep entity resolution with transfer and active learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5851–5861, Florence, Italy. Association for Computational Linguistics.

Kirsch, A., Rainforth, T., and Gal, Y. (2021). Test distribution-aware active learning: A principled approach against distribution shift and outliers.

Kirsch, A., van Amersfoort, J., and Gal, Y. (2019). BatchBALD: Efficient and diverse batch acquisition for deep bayesian active learning. In *Neural Information Processing Systems*, pages 7026–7037.

Konyushkova, K., Sznitman, R., and Fua, P. (2017). Learning active learning from data. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*.

Köksal, A., Schick, T., and Schütze, H. (2023). Meal: Stable and active learning for few-shot prompting.

Le, L., Zhao, G., Zhang, X., Zuccon, G., and Demartini, G. (2024). Colal: Co-learning active learning for text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 13337–13345.

Lewis, D. D. and Gale, W. A. (1994). A sequential algorithm for training text classifiers. In *In Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Li, D., Wang, Z., Chen, Y., Jiang, R., Ding, W., and Okumura, M. (2024). A survey on deep active learning: Recent advances and new frontiers. *arXiv preprint arXiv:2405.00334*.

Li, X. and Qiu, X. (2023). Finding supporting examples for in-context learning.

Liu, J., Shen, D., Zhang, Y., Dolan, B., Carin, L., and Chen, W. (2022). What makes good in-context examples for GPT-3? In Agirre, E., Apidianaki, M., and Vulić, I.,

editors, *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Liu, M., Buntine, W., and Haffari, G. (2018). Learning how to actively learn: A deep imitation learning approach. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1874–1883, Melbourne, Australia. Association for Computational Linguistics.

Maekawa, S., Zhang, D., Kim, H., Rahman, S., and Hruschka, E. (2022). Low-resource interactive active labeling for fine-tuning language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3230–3242.

Margatina, K. and Aletras, N. (2023). On the limitations of simulating active learning. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4402–4419, Toronto, Canada. Association for Computational Linguistics.

Margatina, K., Barrault, L., and Aletras, N. (2022). On the importance of effectively adapting pretrained language models for active learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 825–836, Dublin, Ireland. Association for Computational Linguistics.

Margatina, K., Schick, T., Aletras, N., and Dwivedi-Yu, J. (2023). Active learning principles for in-context learning with large language models. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5011–5034, Singapore. Association for Computational Linguistics.

Margatina, K., Vernikos, G., Barrault, L., and Aletras, N. (2021). Active learning by acquiring contrastive examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 650–663, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

McCallum, A. and Nigam, K. (1998). Employing em and pool-based active learning for text classification. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, page 350–358, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Mehlin, V., Schacht, S., and Lanquillon, C. (2023). Towards energy-efficient deep learning: An overview of energy-efficient approaches along the deep learning lifecycle. *arXiv preprint arXiv:2303.01980*.

Miura, A., Neubig, G., Paul, M., and Nakamura, S. (2016). Selecting syntactic, non-redundant segments in active learning for machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 20–29, Stroudsburg, PA, USA. Association for Computational Linguistics.

Muldrew, W., Hayes, P., Zhang, M., and Barber, D. (2024). Active preference learning for large language models.

Nachtegael, C., De Stefani, J., and Lenaerts, T. (2023). A study of deep active learning methods to reduce labelling efforts in biomedical relation extraction. *PLOS ONE*, 18(12):1–23.

Ni, A., Yin, P., and Neubig, G. (2019). Merging weak and active supervision for semantic parsing. In *AAAI Conference on Artificial Intelligence*.

OpenAI (2023). Gpt-4 technical report.

Pang, J.-C., Fan, H.-B., Wang, P., Xiao, J.-H., Tang, N., Yang, S.-H., Jia, C., Huang, S.-J., and Yu, Y. (2024). Empowering language models with active inquiry for deeper understanding. *arXiv preprint arXiv:2402.03719*.

Pecher, B., Srba, I., Bielikova, M., and Vanschoren, J. (2024). Automatic combination of sample selection strategies for few-shot learning. *arXiv preprint arXiv:2402.03038*.

Qian, K., Chozhiyath Raman, P., Li, Y., and Popa, L. (2020). Learning structured representations of entity names using Active Learning and weak supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6376–6383, Online. Association for Computational Linguistics.

Qian, K., Popa, L., and Sen, P. (2017). Active learning for Large-Scale entity resolution. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1379–1388. ACM.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners.

Rainforth, T., Foster, A., Ivanova, D. R., and Smith, F. B. (2024). Modern Bayesian Experimental Design. *Statistical Science*, 39(1):100 – 114.

Rauch, L., Aßenmacher, M., Huseljic, D., Wirth, M., Bischl, B., and Sick, B. (2023). Activeglae: A benchmark for deep active learning with transformers. In *Machine Learning and Knowledge Discovery in Databases: Research Track*, pages 55–74.

Romberg, J. and Escher, T. (2022). Automated topic categorisation of citizens' contributions: Reducing manual labelling efforts through active learning. In Janssen, M., Csáki, C., Lindgren, I., Loukis, E., Melin, U., Viale Pereira, G., Rodríguez Bolívar, M. P., and Tambouris, E., editors, *Electronic Government*, pages 369–385.

Rouzegar, H. and Makrehchi, M. (2024). Enhancing text classification through llm-driven active learning and human annotation. In *Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 98–111.

Roy, N. and Mccallum, A. (2001). Toward optimal active learning through monte carlo estimation of error reduction. In *in Proceedings of the International Conference on Machine Learning*.

Schröder, C., Müller, L., Niekler, A., and Potthast, M. (2023). Small-text: Active learning for text classification in python. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 84–95, Dubrovnik, Croatia. Association for Computational Linguistics.

Schröder, C. and Niekler, A. (2020). A survey of active learning for text classification using deep neural networks. *CoRR*, abs/2008.07267.

Schröder, C., Niekler, A., and Potthast, M. (2022). Revisiting uncertainty-based query strategies for active learning with transformers. In *Findings of the Association for*

Computational Linguistics: ACL 2022, pages 2194–2203, Dublin, Ireland. Association for Computational Linguistics.

Sener, O. and Savarese, S. (2018). Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*.

Settles, B. (2009). Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Settles, B. and Craven, M. (2008). An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1070–1079, Honolulu, Hawaii. Association for Computational Linguistics.

Settles, B., Craven, M., and Ray, S. (2007). Multiple-instance active learning. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc.

Seung, H. S., Opper, M., and Sompolinsky, H. (1992). Query by committee. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, page 287–294, New York, NY, USA. Association for Computing Machinery.

Shen, Y., Yun, H., Lipton, Z., Kronrod, Y., and Anandkumar, A. (2017). Deep active learning for named entity recognition. In *Proceedings of the Workshop on Representation Learning for NLP*, pages 252–256.

Shi, Z. and Lipani, A. (2023). Don't stop pretraining? make prompt-based fine-tuning powerful learner. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 5827–5849. Curran Associates, Inc.

Shi, Z., Tonolini, F., Aletras, N., Yilmaz, E., Kazai, G., and Jiao, Y. (2023). Rethinking semi-supervised learning with language models. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5614–5634, Toronto, Canada. Association for Computational Linguistics.

Siddhant, A. and Lipton, Z. C. (2018). Deep Bayesian active learning for natural language processing: Results of a large-scale empirical study. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2904–2909, Brussels, Belgium. Association for Computational Linguistics.

Snijders, A., Kiela, D., and Margatina, K. (2023). Investigating multi-source active learning for natural language inference. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2187–2209, Dubrovnik, Croatia. Association for Computational Linguistics.

Steegh, E. and Sileno, G. (2023). No labels? no problem! experiments with active learning strategies for multi-class classification in imbalanced low-resource settings. In *Proceedings of the International Conference on Artificial Intelligence and Law*, ICAIL '23, page 277–286.

Su, H., Kasai, J., Wu, C. H., Shi, W., Wang, T., Xin, J., Zhang, R., Ostendorf, M., Zettlemoyer, L., Smith, N. A., and Yu, T. (2023). Selective annotation makes language models better few-shot learners. In *International Conference on Learning Representations*.

Tamkin, A. (2023). *Foundation Models for the Real World*. PhD thesis. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2024-03-20.

Tamkin, A., Nguyen, D. P., Deshpande, S., Mu, J., and Goodman, N. (2022). Active learning helps pretrained models learn the intended task. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems*.

Tan, Z., Beigi, A., Wang, S., Guo, R., Bhattacharjee, A., Jiang, B., Karami, M., Li, J., Cheng, L., and Liu, H. (2024). Large language models for data annotation: A survey. *arXiv preprint arXiv:2402.13446*.

Tomanek, K. and Hahn, U. (2009). Semi-supervised active learning for sequence labeling. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1039–1047, Suntec, Singapore. Association for Computational Linguistics.

Tomanek, K. and Hahn, U. (2010). A comparison of models for cost-sensitive active learning. In *Coling 2010: Posters*, pages 1247–1255, Beijing, China. Coling 2010 Organizing Committee.

Treviso, M., Lee, J.-U., Ji, T., Aken, B. v., Cao, Q., Ciosici, M. R., Hassid, M., Heafield, K., Hooker, S., Raffel, C., Martins, P. H., Martins, A. F. T., Forde, J. Z., Milder, P., Simpson, E., Slonim, N., Dodge, J., Strubell, E., Balasubramanian, N., Derczynski, L., Gurevych, I., and Schwartz, R. (2023). Efficient Methods for Natural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics*, 11:826–860.

Tsvigun, A., Sanochkin, L., Larionov, D., Kuzmin, G., Vazhentsev, A., Lazichny, I., Khromov, N., Kireev, D., Rubashevskii, A., Shahmatova, O., Dylov, D. V., Galitskiy, I., and Shelmanov, A. (2022). ALToolbox: A set of tools for active learning annotation of natural language texts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 406–434.

Wan, T., Xu, K., Yu, T., Wang, X., Feng, D., Ding, B., and Wang, H. (2023). A survey of deep active learning for foundation models. *Intelligent Computing*, 2:0058.

Wan, Z., Wang, Z., Wang, Y., Wang, Z., Zhu, H., and Satoh, S. (2024). Contributing dimension structure of deep feature for coreset selection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(8):9080–9088.

Wang, X. (2024). Active learning for nlp with large language models. *arXiv preprint arXiv:2401.07367*.

Wang, Z., Zhang, G., Yang, K., Shi, N., Zhou, W., Hao, S., Xiong, G., Li, Y., Sim, M. Y., Chen, X., et al. (2023). Interactive natural language processing. *arXiv preprint arXiv:2305.13246*.

Wei, Q., Chen, Y., Salimi, M., Denny, J. C., Mei, Q., Lasko, T. A., Chen, Q., Wu, S., Franklin, A., Cohen, T., and Xu, H. (2019). Cost-aware active learning for named entity recognition in clinical text. *Journal of the American Medical Informatics Association*, 26(11):1314–1322.

Ying, Y., Lidong, Z., and Changjie, C. (2024). A support data-based core-set selection method for signal recognition. *China Communications*, 21(4):151–162.

Yu, Y., Kong, L., Zhang, J., Zhang, R., and Zhang, C. (2022). AcTune: Uncertainty-based active self-training for active fine-tuning of pretrained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1422–1436, Seattle, United States. Association for Computational Linguistics.

Yu, Y., Zhang, R., Xu, R., Zhang, J., Shen, J., and Zhang, C. (2023). Cold-start data selection for better few-shot language model fine-tuning: A prompt-based uncertainty propagation approach. In *Annual Meeting of the Association for Computational Linguistics*, pages 2499–2521.

Yuan, M., Xia, P., May, C., Van Durme, B., and Boyd-Graber, J. (2022). Adapting coreference resolution models through active learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7533–7549, Dublin, Ireland. Association for Computational Linguistics.

Zeng, X. and Zubiaga, A. (2023). Active PETs: Active data annotation prioritisation for few-shot claim verification with pattern exploiting training. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 190–204, Dubrovnik, Croatia. Association for Computational Linguistics.

Zhang, M. and Plank, B. (2021). Cartography active learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 395–406, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhang, R., West, R., Cui, X., and Zhang, C. (2022a). Adaptive multi-view rule discovery for weakly-supervised compatible products prediction. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 4521–4529.

Zhang, R., Yu, Y., Shetty, P., Song, L., and Zhang, C. (2022b). Prompt-based rule discovery and boosting for interactive weakly-supervised learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 745–758, Dublin, Ireland. Association for Computational Linguistics.

Zhang, R., Yu, Y., and Zhang, C. (2020). SeqMix: Augmenting active sequence labeling via sequence mixup. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8566–8579, Online. Association for Computational Linguistics.

Zhang, S., Gong, C., Liu, X., He, P., Chen, W., and Zhou, M. (2022c). ALLSH: Active learning guided by local sensitivity and hardness. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1328–1342, Seattle, United States. Association for Computational Linguistics.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., and Zettlemoyer, L. (2022d). Opt: Open pre-trained transformer language models.

Zhang, Z. (2023). *Exploring Language Structured Prediction in Resource-limited Scenarios*. PhD thesis, Carnegie Mellon University.

Zhang, Z., Strubell, E., and Hovy, E. (2022e). A survey of active learning for natural language processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Zhang, Z., Strubell, E., and Hovy, E. (2022f). A survey of active learning for natural language processing. In *Conference on Empirical Methods in Natural Language Processing*, pages 6166–6190.

Zhao, Y., Zhang, H., Zhou, S., and Zhang, Z. (2020). Active learning approaches to enhancing neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1796–1806, Online. Association for Computational Linguistics.

Zhdanov, F. (2019). Diverse mini-batch active learning.

Zheng, H., Guan, M., Mei, Y., Li, Y., and Wu, Y. (2023). Ecnu-llm@ chip-promptcblue: Prompt optimization and in-context learning for chinese medical tasks. In *China Health Information Processing Conference*, pages 60–72. Springer.

Zhu, J., Wang, H., Yao, T., and Tsou, B. K. (2008). Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 1137–1144, Manchester, UK. Coling 2008 Organizing Committee.