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.

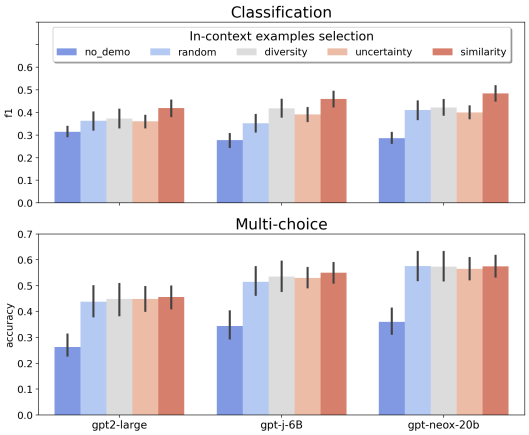


Figure 1: Performance of different in-context selection algorithms in classification and multi-choice tasks.

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

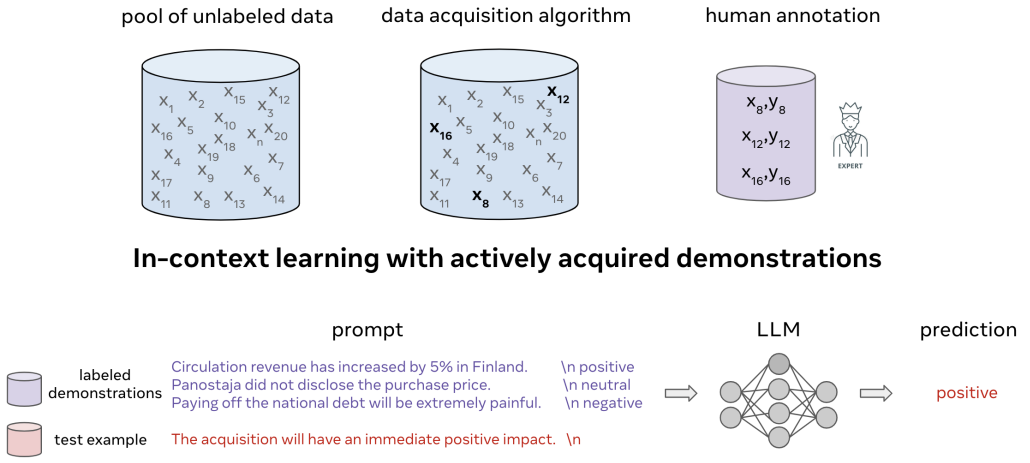


Figure 2: Top: Active data collection (single iteration). Bottom: Prompt construction and model inference.

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.

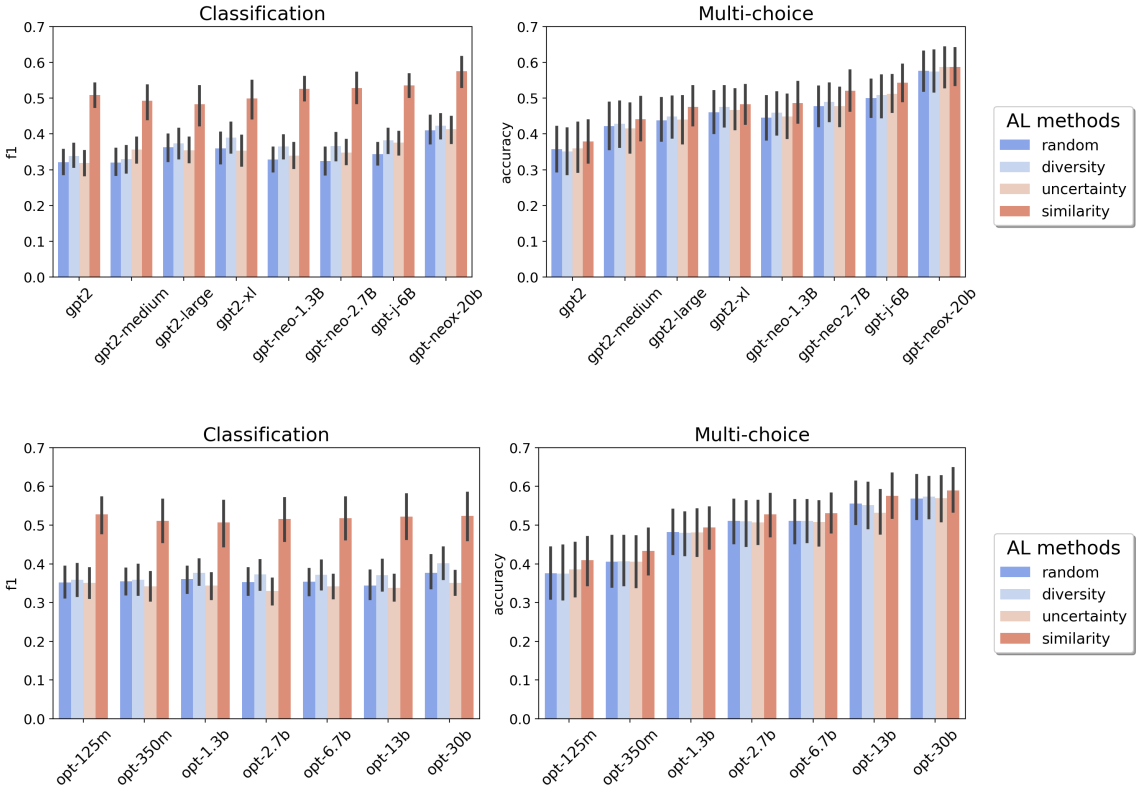


Figure 3: Results for various GPT (top) and OPT (bottom) models and AL methods averaged over 15 classification and 9 multi-choice tasks. *Similarity* is consistently the best performing approach overall, followed by *diversity* and *random*. Interestingly, we observe that *uncertainty* sampling underperforms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

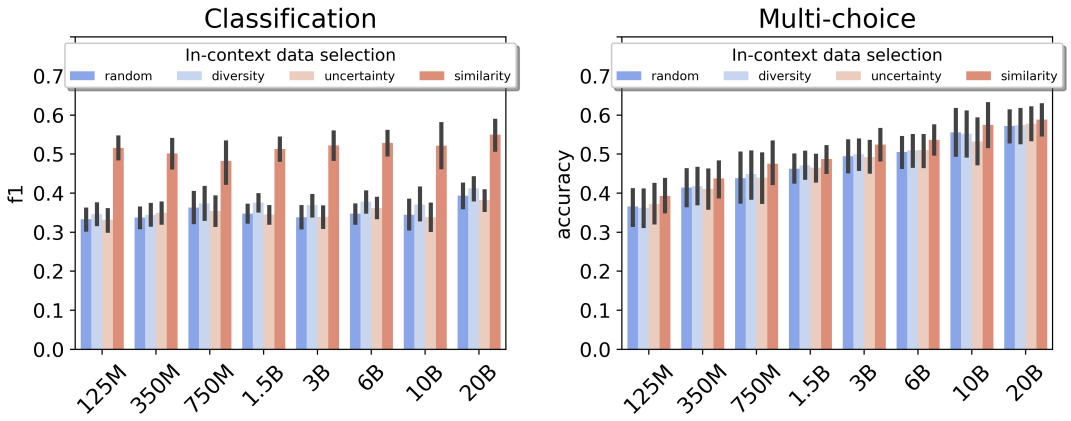


Figure 4: Results per model size.

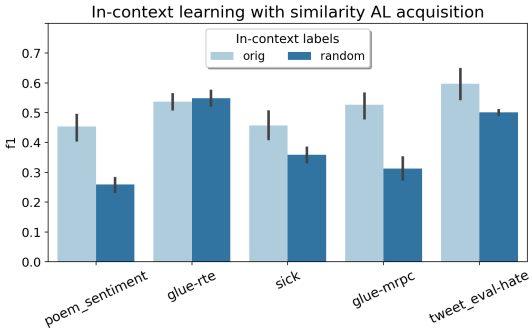


Figure 5: Effect of ground truth labels on in-context learning with with the similarity AL selection method.

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-

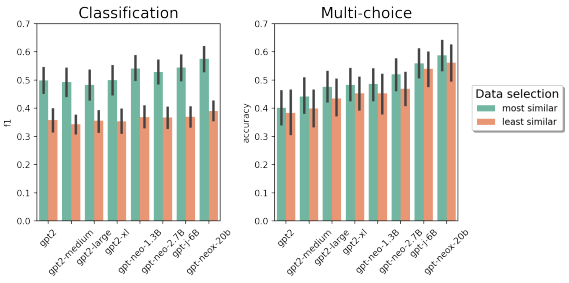


Figure 6: Most vs. least similar in-context examples.

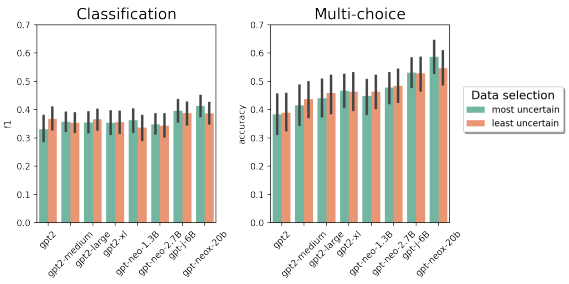


Figure 7: Most vs. least uncertain in-context examples.

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-

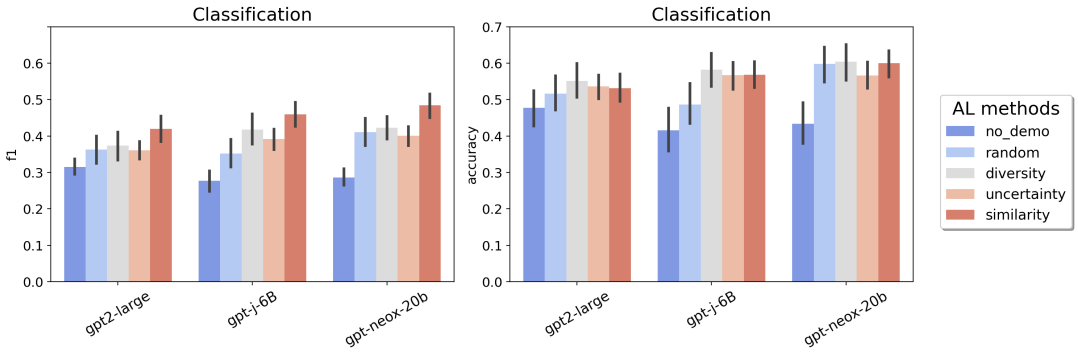


Figure 8: The ranking of data selection methods is different depending on the metric used.

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.

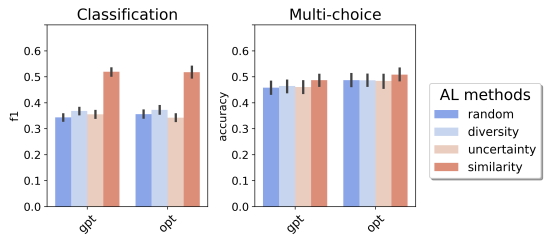


Figure 9: Results per model family.

A Experimental Details

A.1 Tasks & Datasets

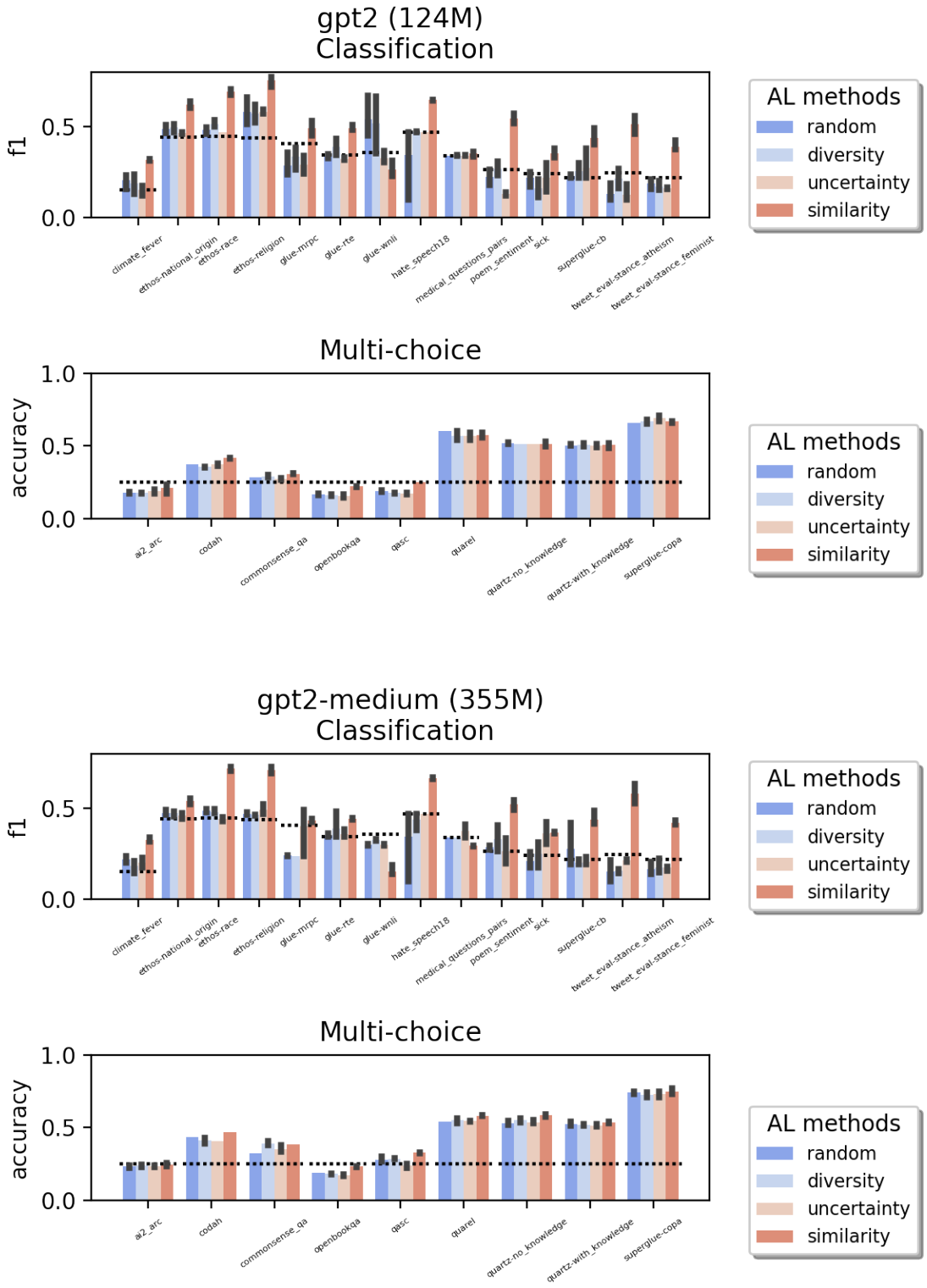
Following Min et al. (2022), we evaluate our models in 15 classification and 9 multi-choice tasks taken from the `Crossfit` (Ye et al., 2021) benchmark. Specifically the tasks we evaluate are *poem_sentiment* (Sheng and Uthus, 2020), *glue-wnli* (Wang et al., 2019; Levesque et al., 2012), *climate_fever* (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.

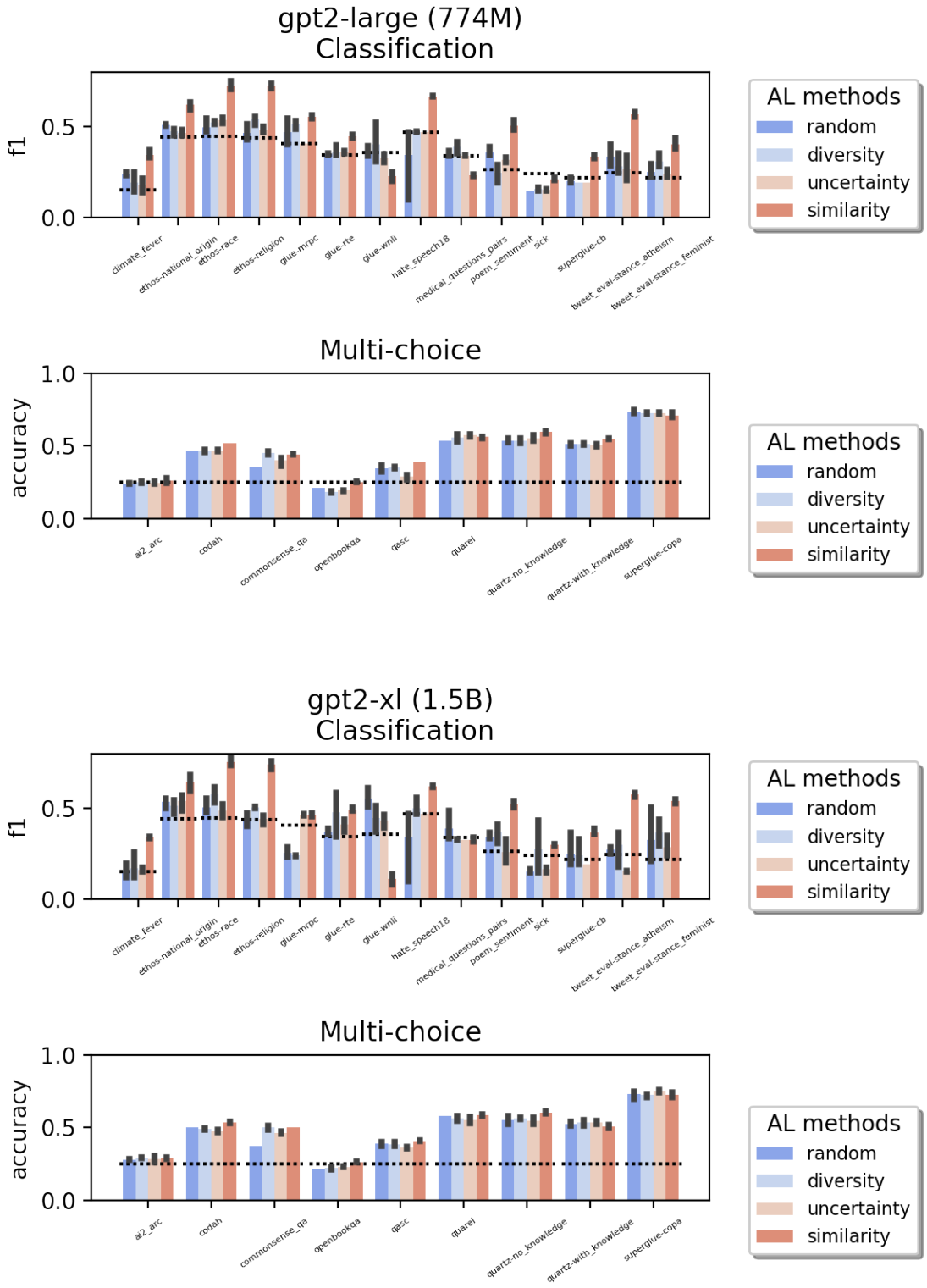
A.2 Full results

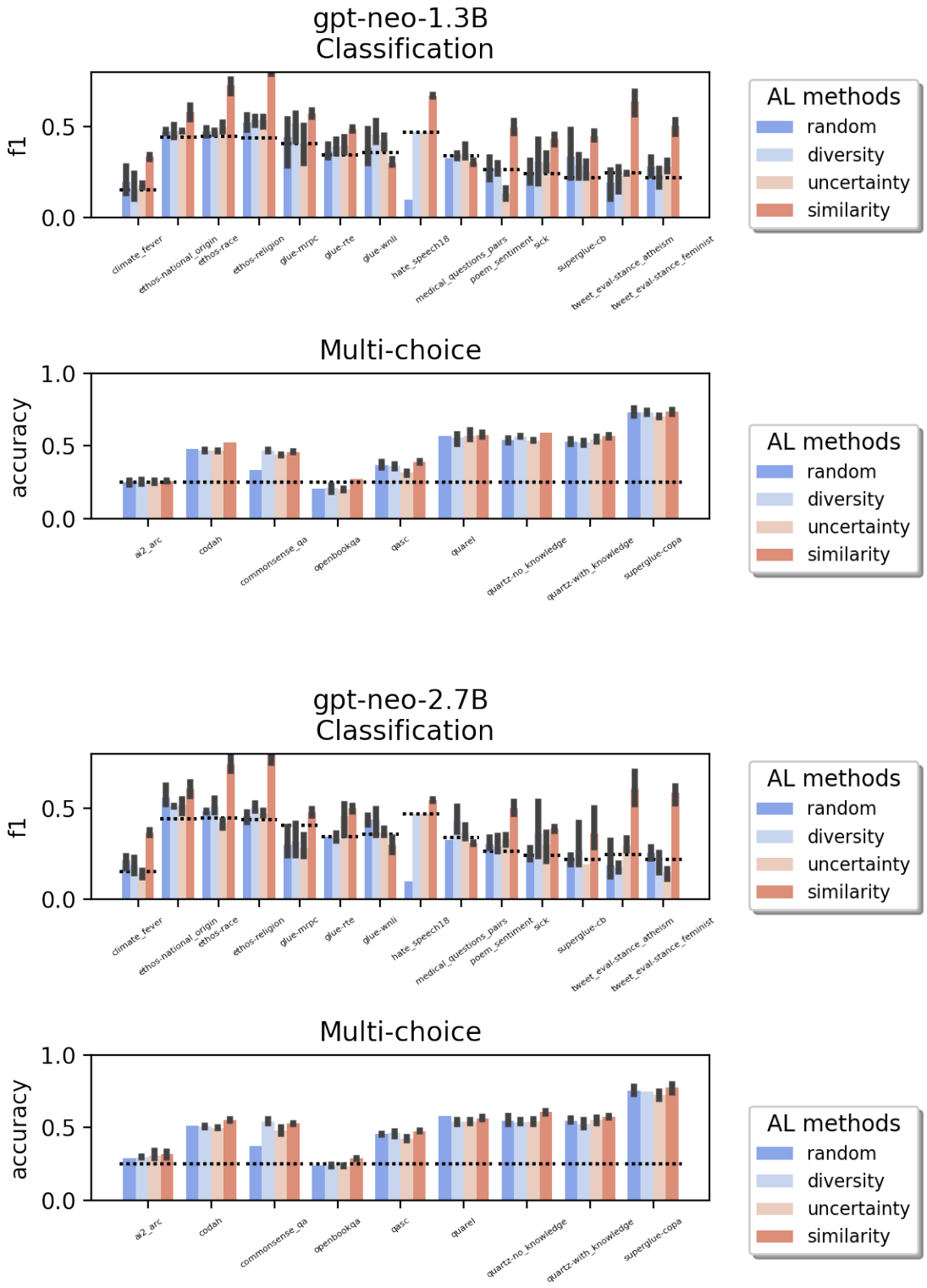
We provide below the full set of results, for each dataset, model and active learning acquisition strategy considered. The dashed line depicts the majority vote baseline.

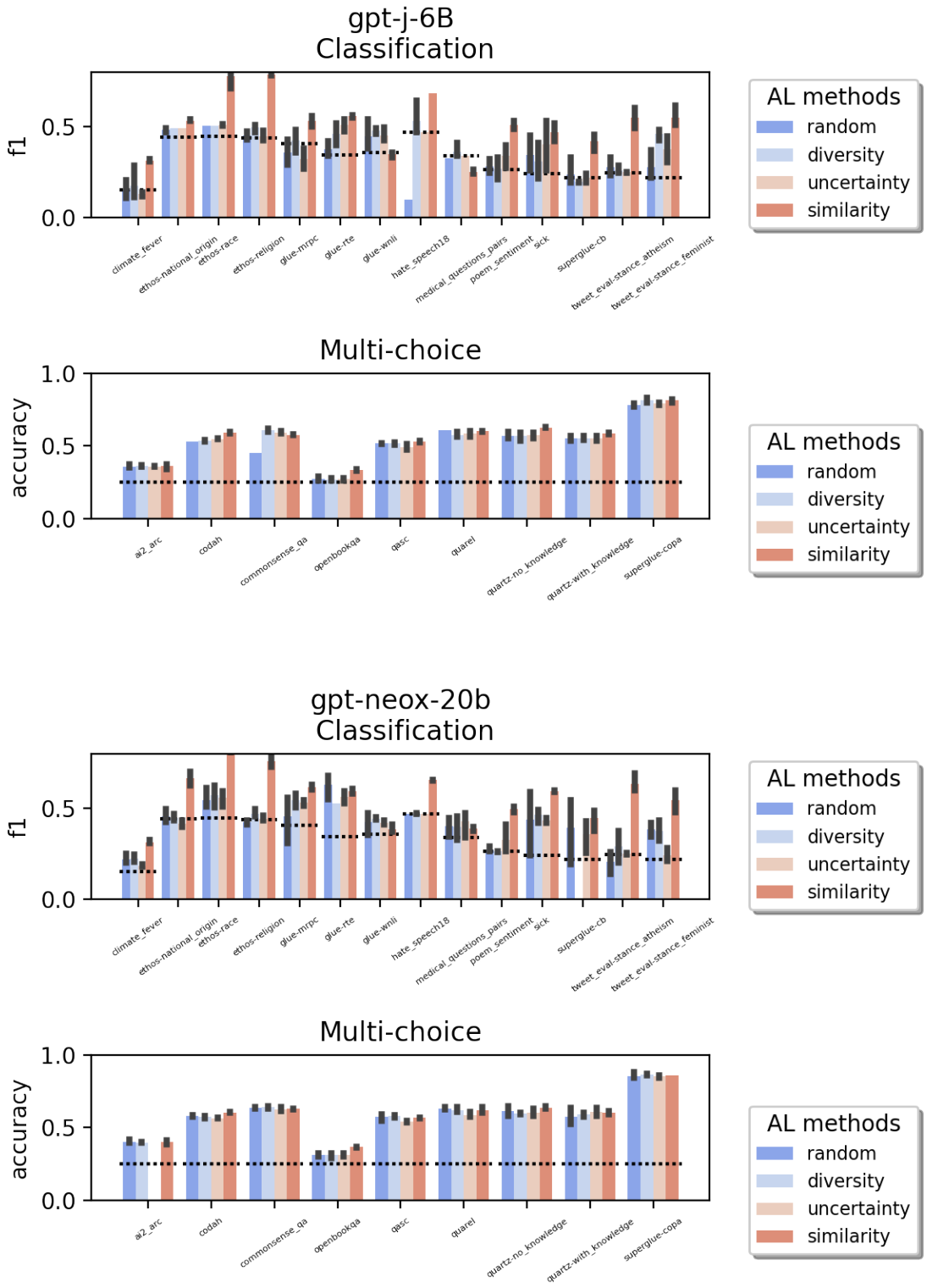
A.3 Model Family

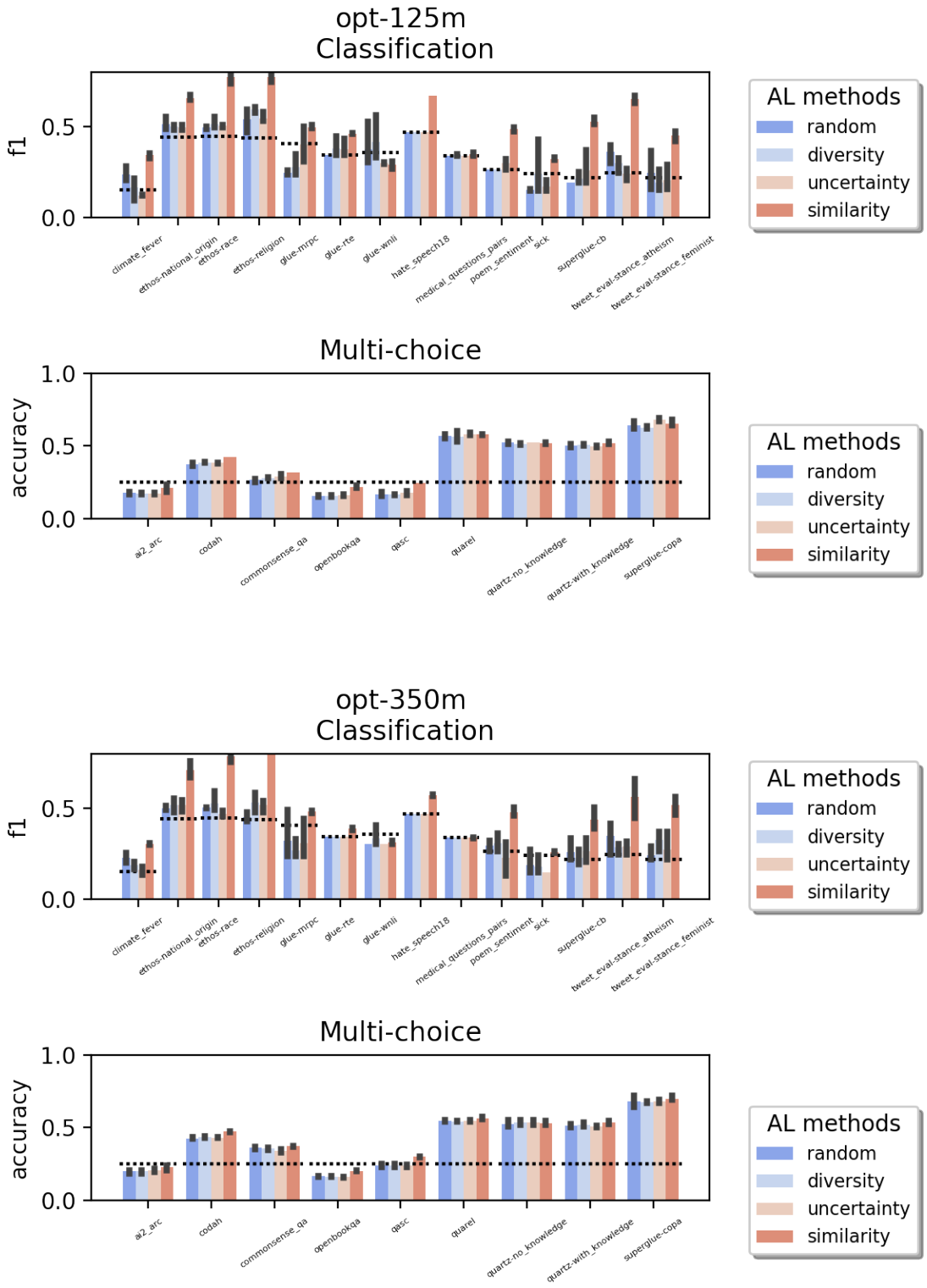
We provide the results on few-shot learning with k=16 demonstrations per prompt per model family and task type in Figure 9. We observe the same patterns for both GPT and OPT models.

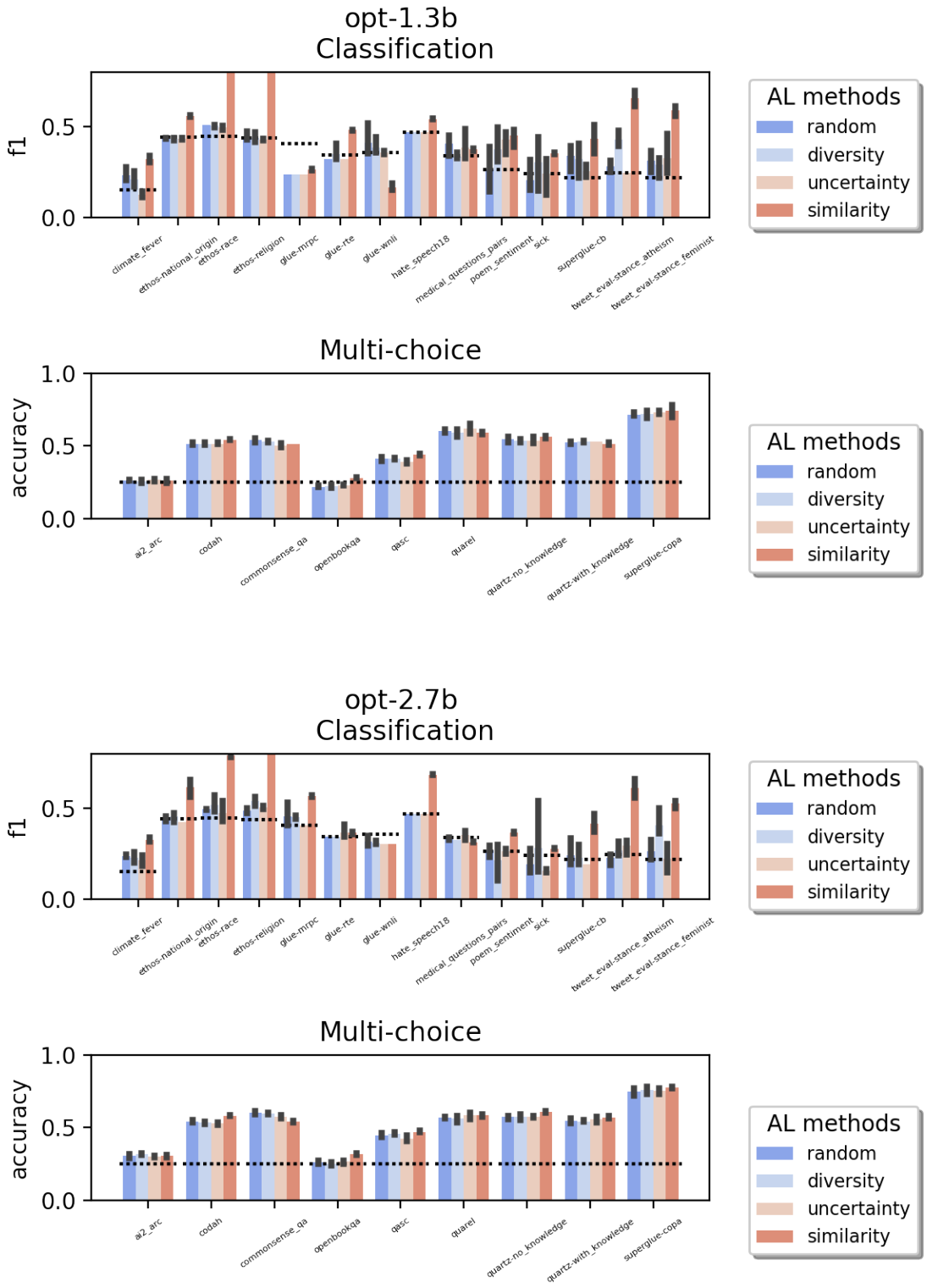

gpt2 (124M)
Classification
f1
0.0
0.5
climate_fever
ethos-national_origin
ethos-race
ethos-religion
glue-mrpc
glue-rte
glue-wnli
hate_speech18
medical_questions_pairs
poem_sentiment
sick
superglue-cb
tweet_eval-stance_atheism
tweet_eval-stance_feminist
AL methods
random
diversity
uncertainty
similarity
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

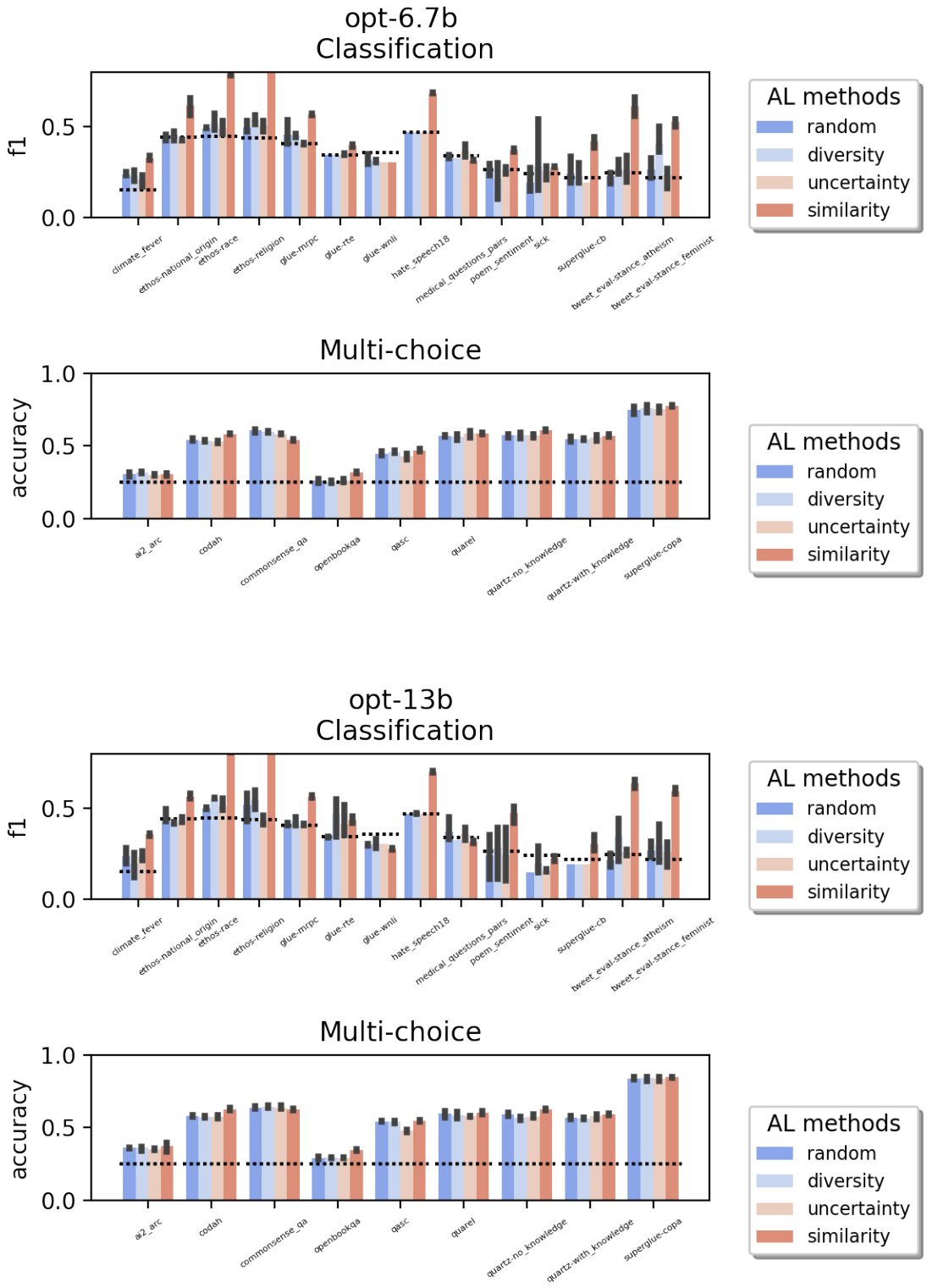

gpt2-large (774M)
Classification
f1
0.0
0.5
climate_fever
ethos-national_origin
ethos-race
ethos-religion
glue-mrpc
glue-rte
glue-wnli
hate_speech18
medical_questions_pairs
poem_sentiment
sick
superglue-cb
tweet_eval-stance_atheism
tweet_eval-stance_feminist
AL methods
random
diversity
uncertainty
similarity
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

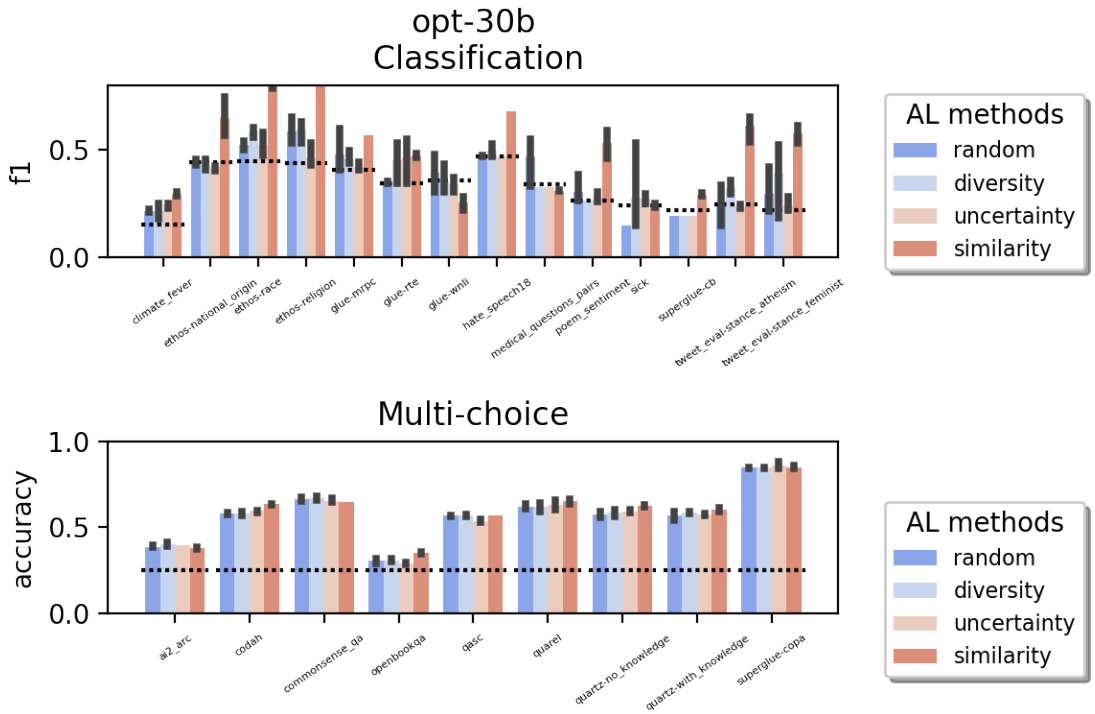

gpt-neo-1.3B
Classification
f1
0.0
0.5
climate_fever
ethos-national_origin
ethos-race
ethos-religion
glue-mrpc
glue-rte
glue-wnli
hate_speech18
medical_questions_pairs
poem_sentiment
sick
superglue-cb
tweet_eval-stance_atheism
tweet_eval-stance_feminist
AL methods
random
diversity
uncertainty
similarity
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