How many examples does it take for fine-tuning to outperform few-shot prompting? A study of medical text classification and domain adaptation

Anonymous ACL submission

Abstract

Given the recent success of large language models, a critical question for machine learning engineers is when to use few-shot prompting vs. fine-tuning. We explore this question in a medical setting, where data restrictions make only a small number of training examples realistic, and where the ability to adapt from one domain to another is critical. On two medical text classification tasks, we find that fine-tuning outperforms few-shot prompting with as little as 100 labeled examples and that few-shot prompting has a greater risk of robustness problems.

1 Introduction

Adapting NLP models to new tasks and domains in the medical field is challenging. Patient privacy constraints severely limit the sharing of large, annotated datasets across institutions. If data sharing is permitted, it often requires complex agreements, resource-intensive de-identification processes, and expert annotation to ensure no protected health information (PHI) is disclosed. Under these conditions, large annotated corpora for fine-tuning models are infeasible, highlighting the need for methods that can work with very limited data.

Large language models (LLMs) have shown remarkable abilities in few-shot generalization, effectively leveraging a handful of labeled examples to perform new tasks. The computational cost of LLM inference can be substantial, but smaller models can require more data to achieve similar performance. This trade-off raises a practical question: *Given a limited annotation budget, should we invest in fully fine-tuning a smaller model or leverage few-shot prompting of a large model?*

To address this question, we consider both traditional fine-tuning and few-shot prompting approaches. Our analysis is guided by the following research questions:

RQ1: *For a given task and a fixed number of labeled samples, which approach yields better performance, fine-tuning or few-shot prompting?*

RQ2: *When using few-shot prompting, what model size and how many examples are needed to achieve reasonable performance?*

RQ3: *When using fine-tuning, what model size and how many examples are needed to achieve reasonable performance?*

RQ4: *When transferring a model to a new domain, which approach is more robust, fine-tuning or few-shot prompting?*

We investigate these research questions by few-shot prompting and fine-tuning language models of various sizes on medical text classification tasks over various domains. Our main contributions are:

- We determine the cross-over point where labeled examples are better used for fine-tuning than for few-shot prompting. For our medical NLP tasks, with only 40 labeled examples, fine-tuning Llama 3.2 3B is better than prompting it. With only 160 labeled examples, fine-tuning the tiny RoBERTa is better than prompting the huge Llama 3.1 70B.
- We find that few-shot prompted models are not consistently more robust than fine-tuned models on new domains. For our medical NLP tasks, few-shot prompted models are slightly more robust on causal classification, but much less robust on negation classification.

2 Related work

Both few-shot prompting and fine-tuning of LLMs yields strong performance across a variety of NLP tasks, including translation, question answering, and text classification. Few-shot prompting has shown impressive results in tasks ranging from machine translation and question answering to tabular data classification and relation extraction (Brown et al., 2020; Xu et al., 2023; Hegselmann et al., 2023; Ma et al., 2023; Touvron et al., 2023). Con-

currently, fine-tuning LLMs has proven effective not only for machine translation and classification (Zhang et al., 2023; Hsieh et al., 2023; Edwards and Camacho-Collados, 2024), but also for a broad array of benchmark tasks (Chung et al., 2024).

In the medical domain, studies have investigated the domain adaptation capabilities of LLMs via prompting and fine-tuning (Van Veen et al., 2023; Fan et al., 2023; Labrak et al., 2024). For example, Van Veen et al. (2023) explores adaptation strategies for domain shifts in radiology reports. Research outside the medical field has developed benchmarks to evaluate how fine-tuned and few-shot prompted models withstand shifts across various domains (Calderon et al., 2024). Recent work has also examined introducing extra parameters to improve LLMs' resilience to domain shifts (Huang et al., 2023; Ormazabal et al., 2023).

However, prior efforts have not systematically investigated how performance compares as a function of available labeled data. They do not pinpoint the exact threshold at which fine-tuned models consistently outperform their few-shot prompted counterparts. Our study fills this gap by examining a range of data scenarios, offering practical guidance on whether to invest in full fine-tuning or few-shot prompting given a specific annotation budget.

3 Tasks and Datasets

We focus on two medical text classification tasks: causal classification and negation classification. Each includes two datasets from distinct domains, allowing us to study domain adaptation settings.

Causal Classification. Yu et al. (2020) aims to detect when correlational findings are overstated as causal claims. It classifies claim sentences in press releases and PubMed articles into four categories: *Correlational*, *Conditional causal*, *Direct causal*, or *Not claim*. For instance, *"Suicide risk greater for people living at higher elevations"* is labeled as *Correlational*, while *"Traffic noise increases the risk of having a stroke."* is *Direct causal*. We use press releases as the source domain and PubMed as the target domain, thus exploring adaptation from public-facing summaries to scientific abstracts.

Negation Classification. Derived from SemEval 2021 Task 10 (Laparra et al., 2021), this task identifies whether a medical event (marked in the sentence) is negated by its context. For example, in *"Has no <e> diarrhea </e> and no new lumps or masses"*, the event *diarrhea* should be classified as negated. We use i2b2 (Partners HealthCare) as the source domain and MIMIC (Beth Israel ICU notes) as the target domain, thus exploring adaptation between two distinct institutions.

We select these two tasks and four datasets for:

1. **Simplicity:** Both tasks involve straightforward classification, minimizing the complexity of reasoning chains or prompt engineering. This allows for more direct comparisons between fine-tuning and few-shot prompting.
2. **Medical Domain Relevance:** Operating in the medical domain aligns with practical constraints on data sharing and annotation. The difficulty of exchanging patient-related data underscores the importance of domain adaptation techniques.
3. **Challenging Domain Transfer:** Our preliminary results on the negation classification task, along with previous works (Laparra et al., 2021; Su et al., 2022) show performance degradations when models trained on one medical subdomain are applied to another.

See Appendix A for additional dataset and evaluation details.

4 Models and Approaches

Models. We consider two categories of models:

- **Large generative:** We use the open-source LLaMA LLMs (Touvron et al., 2023) at varying scales: `Llama-3.2-3B`, `Llama-3.1-8B`, and `Llama-3.1-70B`.
- **Small encoder-only:** We use the open-source RoBERTa model (Liu, 2019), `roberta-base`.

Including both categories allows us to contrast full fine-tuning of a smaller encoder-only model with few-shot prompting of a larger generative model.

Few-Shot Prompting. For few-shot prompting, we adopt a simple approach. We prepend the specified number of labeled source-domain input-output pairs before the test instance, allowing the LLM to infer patterns from these exemplars. This approach avoids extensive prompt engineering, providing a clear baseline for few-shot adaptation.

Fine-Tuning. For fine-tuning, we adopt the standard HuggingFace `Trainer` API (Wolf et al., 2020), fine-tuning all model parameters on the specified number of labeled source-domain input-output pairs. To remain practical, we limit fine-tuning to *RoBERTa-base* and *LLaMA-3.2-3B*, given the computational overhead of fine-tuning larger models.

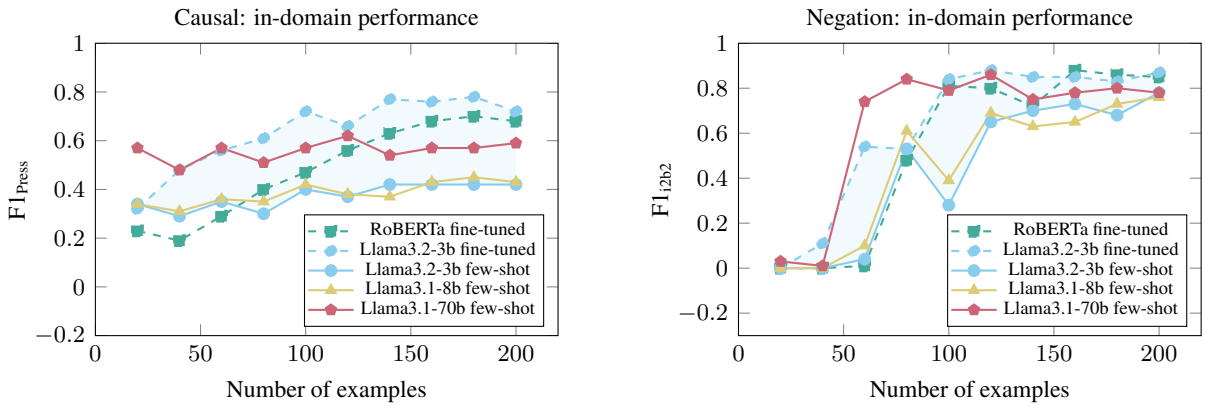

Figure 1: Fine-tuning (dashed lines) outperforms few shot prompting (solid lines) with only a few examples, for both causal classification and negation classification. The shaded area shows the fine-tuning vs. few-shot-prompting difference for the comparable Llama-3.3-2b models. Llama-3.3-2b when fine-tuned outperforms Llama-3.3-2b when used few-shot with 40 or more examples. Even the tiny RoBERTa when fine-tuned outperforms the huge Llama3.1-70b used few-shot with 160 or more examples.

- **RoBERTa:** We treat the problem as sequence classification, adding a special <s> token at the input's start and classifying the entire sentence based on its representation.
- **LLaMA-3.2-3B:** We use a causal language modeling approach, training the model to generate the class label tokens given the input. The input tokens are masked from the loss calculation.

5 Experimental Design and Data Budget

To simulate realistic conditions, we assume that the amount of data available from the source domain is limited and must be thoroughly vetted for PHI. We therefore incrementally vary the number of labeled samples from 20 to 200 in steps of 20. At each step, we compare fine-tuning and few-shot prompting to examine how each method scales with increasing yet still modest annotation budgets.

This incremental approach enables a systematic exploration of the trade-offs between data availability, annotation cost, computational expense, and final model quality. By focusing on small yet realistic data budgets, we offer insights that are relevant to practical medical NLP scenarios where data scarcity and privacy constraints are the norm.

6 Results

RQ1: Fine-Tuning vs. Few-Shot Prompting

Figure 1 shows that as the number of labeled samples increases even modestly, fine-tuned models outperform larger models that rely on few-shot prompting. For instance, in causal classification, fine-tuning a small model (the 125M parameter RoBERTa-base) on about 140 samples surpasses the few-shot performance of LLaMA models with 3B, 8B, or even 70B parameters. Similarly, in negation classification, a fine-tuned RoBERTa-base model trained on roughly 160 samples outperforms few-shot prompted LLaMA models of all sizes.

Few-shot prompting offers advantages under extreme data scarcity with very large models. For causal classification, having only 20 samples favors a few-shot 70B-parameter LLaMA model. For negation classification, having 80 or fewer samples favors the 70B model. This advantage disappears once the available data crosses a minimal threshold (40–100 samples) and fine-tuning a small model becomes the more effective option.

RQ2: Few-Shot Prompting: Model Size and Number of Examples

Few-shot prompting is a competitive option when a large model is available and the number of samples is extremely limited. The best few-shot prompting model that outperforms fine-tuning uses only 20 samples for causal classification and 80 samples for negation classification. In both cases, only the large 70B-parameter LLaMA model outperforms fine-tuning; smaller LLaMA models fail to exhibit this advantage. This aligns with prior work (Touvron et al., 2023) that links stronger few-shot performance to larger model sizes.

RQ3: Fine-Tuning: Model Size and Number of Examples

Fine-tuning is a competitive option whenever a modest number of samples (40-100) is available.

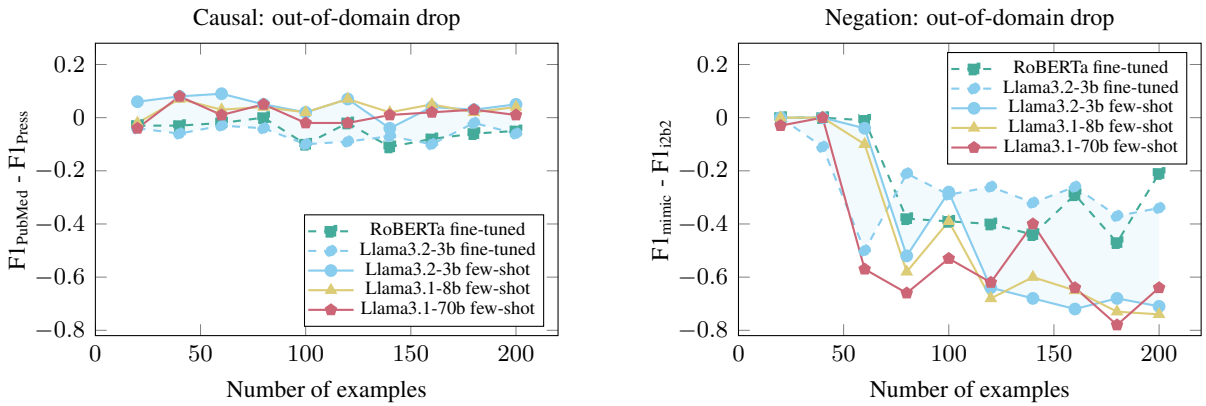

Figure 2: For causal classification, few-shot prompting (solid lines) is slightly more robust to changes in domain than fine-tuning (dashed lines), though neither approach sees large performance drops when models are trained/prompted with examples from press releases but tested on PubMed. For negation classification, fine-tuning (dashed lines) shows much smaller drops in performance than few-shot prompting (solid lines) when models are trained/prompted with examples from i2b2 but tested on MIMIC. The shaded area shows the fine-tuning vs. few-shot-prompting difference for the comparable Llama-3.3-2b models.

Both the causal and negation classification tasks show steady gains as the number of labeled examples increases, regardless of model size. Though larger models outperform smaller models, these gains are modest compared to the gains from increasing the dataset size, and mostly disappear by 200 samples. For example, in negation classification, once the count reaches 160 examples, the performance gap between a fine-tuned LLaMA-3.2-3B and a fine-tuned RoBERTa-base disappears. Thus, focusing on data quality and quantity yields greater returns than scaling up model size.

RQ4: Domain Adaptation: Fine-Tuning or Few-Shot Prompting?

The left half of Figure 2 shows that in causal classification, few-shot prompting demonstrates slightly more robust domain generalization. The average performance change for models given training examples from press releases and tested on examples from PubMed is +0.03 for few-shot prompted models and -0.05 for fine-tuned models. Few-shot prompting is slightly better, but neither approach experiences catastrophic degradation, suggesting that both methods are mostly resilient when shifting from press releases to PubMed.

The right half of Figure 2 shows that in negation classification, both approaches see significant degradation, with few-shot prompted models being much less robust. The average performance change for models given training examples from i2b2 data and tested on MIMIC notes is -0.26 for fine-tuned models and -0.45 for few-shot prompted models. For example, with 200 labeled samples, a fine-tuned RoBERTa-base model incurs a performance drop of about 0.2, while a 70B-parameter LLaMA few-shot model's performance plummets by roughly 0.6. In scenarios where labeled target-domain data is unavailable, fine-tuning a source-domain model appears safer and more stable.

7 Conclusion

Our comparisons of few-shot learning and fine-tuning on two medical text classification tasks across four domains reveal a number of useful findings for machine learning engineers. For tasks and domains like ours, few-shot prompting is viable only in cases of severe data scarcity, where only 20-80 labeled examples are available. When even 100 labeled examples are available, fine-tuning a Llama3.2-3B model yields better performance than using the much larger Llama3.1-70B with a few-shot prompt. Fine-tuning the even smaller RoBERTa model yields similar performance to the larger fine-tuned Llama3.2-3B if as little as 200 labeled examples are available. If domain adaptation is a concern, fine-tuning is the less risky option, as few-shot prompting ranges from slightly more robust across domains (0.08 better than fine-tuning) to much less robust (0.19 worse than fine-tuning). Future work is needed to see how these results generalize across tasks beyond text classification and domains beyond medicine.

Limitations

As this is a short article, the number of tasks, domains, and languages explored was limited: two medical text classification tasks and four domains, all in English. Future work should explore other tasks, such as information extraction or question answering, and other domains, such as social media or legal documents, as well as additional languages.

Exploration of additional tasks is also necessary to understand when few-shot prompting will become less robust as it does in the negation classification task. One explanation of this phenomenon might be the high label imbalance: in the negation classification dataset, the source domain contained 1,115 negations versus 4,430 non-negations, while the target domain contained 958 negations versus 8,622 non-negations. By contrast, the causal classification dataset maintained a more balanced label distribution: in the source domain, it included 738 correlational, 568 direct causal, 486 no claim, and 284 conditional causal instances; in the target domain, 1,356 correlational, 998 direct causal, 494 no claim, and 213 conditional causal instances. But to confidently attribute this difference to label imbalance we would need to find other tasks with similar label imbalance that also show this phenomenon. Another explanation of the phenomenon might be divergence of annotation guidelines, e.g., inherently negated words like *afebrile* are marked as negated in i2b2 but as non-negated in MIMIC. But to confidently attribute this difference to annotation guideline divergence, we would need to find other tasks with similar annotation divergences that also show this phenomenon.

We did not explore strategies for balancing labels, though both fine-tuning or few-shot prompting methods can sometimes benefit from label balancing. This was because we assumed a small annotation budget ($\leq$ 200 examples), and balancing labels requires annotating additional data and then sampling down. So, for example, given that negations constitute roughly 20% of the source domain data, to get 20 negation classification training examples (10 negated and 10 non-negated), we would need to annotate approximately 50 examples in total, from which we would expect around 10 negated and 40 non-negated instances. Future work should explore whether the benefits of label balancing outweigh the additional annotation costs.

We did not extensively engineer the prompts for either the few-shot prompted models or the fine-tuned models. The simplicity of our tasks meant that simple input-output pairs as a prompt worked sufficiently well. But it is possible that both the few-shot prompted models and the fine-tuned models might achieve higher performance with additional prompt engineering.

Ethics Considerations

We use LLMs throughout our experiments. While LLMs can potentially produce harmful or biased content (Bianchi et al., 2024), we limit their usage in this study to generating class labels for classification tasks. This restricted application reduces the likelihood of unintended harmful output.

References

Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio, Paul Rottger, Dan Jurafsky, Tatsunori Hashimoto, and James Zou. 2024. Safety-tuned LLaMAs: Lessons from improving the safety of large language models that follow instructions. In *The Twelfth International Conference on Learning Representations*.

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.

Nitay Calderon, Naveh Porat, Eyal Ben-David, Alexander Chapanin, Zorik Gekhman, Nadav Oved, Vitaly Shalumov, and Roi Reichart. 2024. Measuring the robustness of NLP models to domain shifts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 126–154, Miami, Florida, USA. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Aleksandra Edwards and Jose Camacho-Collados. 2024. Language models for text classification: Is in-context learning enough? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10058–10072, Torino, Italia. ELRA and ICCL.

Longjun Fan, Xiaohong Liu, Yuhao Wang, Guoxing Yang, Zongxin Du, and Guangyu Wang. 2023. Enhancing medical language understanding: Adapting llms to the medical domain through hybrid granularity mask learning. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2990–2995. IEEE.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. 2023. Tabllm: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics*, pages 5549–5581. PMLR.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017, Toronto, Canada. Association for Computational Linguistics.

Yangsibo Huang, Daogao Liu, Zexuan Zhong, Weijia Shi, and Yin Tat Lee. 2023. k nn-adapter: Efficient domain adaptation for black-box language models. *arXiv preprint arXiv:2302.10879*.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*.

Egoitz Laparra, Xin Su, Yiyun Zhao, Özlem Uzuner, Timothy Miller, and Steven Bethard. 2021. SemEval-2021 task 10: Source-free domain adaptation for semantic processing. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 348–356, Online. Association for Computational Linguistics.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Huan Ma, Changqing Zhang, Yatao Bian, Lemao Liu, Zhirui Zhang, Peilin Zhao, Shu Zhang, Huazhu Fu, Qinghua Hu, and Bingzhe Wu. 2023. Fairness-guided few-shot prompting for large language models. *Advances in Neural Information Processing Systems*, 36:43136–43155.

Aitor Ormazabal, Mikel Artetxe, and Eneko Agirre. 2023. CombLM: Adapting black-box language models through small fine-tuned models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2961–2974, Singapore. Association for Computational Linguistics.

Xin Su, Yiyun Zhao, and Steven Bethard. 2022. A comparison of strategies for source-free domain adaptation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8352–8367, Dublin, Ireland. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Dave Van Veen, Cara Van Uden, Maayane Attias, Anuj Pareek, Christian Bluethgen, Malgorzata Polacin, Wah Chiu, Jean-Benoit Delbrouck, Juan Zambrano Chaves, Curtis Langlotz, Akshay Chaudhari, and John Pauly. 2023. RadAdapt: Radiology report summarization via lightweight domain adaptation of large language models. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 449–460, Toronto, Canada. Association for Computational Linguistics.

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 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Xin Xu, Yuqi Zhu, Xiaohan Wang, and Ningyu Zhang. 2023. How to unleash the power of large language models for few-shot relation extraction? In *Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pages 190–200, Toronto, Canada (Hybrid). Association for Computational Linguistics.

Bei Yu, Jun Wang, Lu Guo, and Yingya Li. 2020. Measuring correlation-to-causation exaggeration in press releases. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4860–4872, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Xuan Zhang, Navid Rajabi, Kevin Duh, and Philipp Koehn. 2023. Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with QLoRA. In *Proceedings of the Eighth Conference on Machine Translation*, pages 468–481, Singapore. Association for Computational Linguistics.

A Dataset Details

We use the datasets released by Yu et al. (2020) for the causal classification task. From each domain, we select the first 1,000 samples as the training set and use the remainder as the test set. In the source

domain, the training set contains 1,000 samples and the test set contains 1,076 samples. In the target domain, the training set contains 1,000 samples and the test set contains 2,061 samples.

For the negation classification task, we follow the original SemEval 2021 Task 10 (Laparra et al., 2021) data splits, using their development sets as our training sets and their test sets as our test sets. In the source domain, the training set has 1,109 samples and the test set has 4,436 samples. In the target domain, the training set has 1,916 samples and the test set has 7,664 samples.

All training and few-shot samples are randomly drawn from the source domain training sets. We always evaluate on the corresponding test sets. Following the original evaluation metrics, we report the macro-averaged F_1 score for the causal classification task. For the negation classification task, we report the F_1 score for the negated class.

B License Information

We comply with all relevant model licenses and adhere to the intended uses defined by their creators. The LLaMA 3.1 models are provided under the Llama 3.1 Community License Agreement, while the LLaMA 3.2 models are distributed under the Llama 3.2 Community License Agreement. RoBERTa is released under the MIT License.

We use the HuggingFace Transformers library (Wolf et al., 2020) for fine-tuning, which is licensed under the Apache-2.0 license. For inference with LLMs, we rely on the vLLM library, also licensed under Apache-2.0.

All datasets are used in compliance with their respective licenses. The causal classification data from Yu et al. (2020) are released under the GPL-3.0 license. The negation classification data from SemEval 2021 Task 10 (Laparra et al., 2021) are provided under the Apache-2.0 license.

C Implementation Details

We fine-tune all LLMs using their *instruct* versions and perform few-shot prompting using their *base* versions. Our preliminary experiments show that instruct versions often produce extraneous text in few-shot prompting settings, making performance evaluation more difficult.

We run all experiments on NVIDIA A100 GPUs with 80GB of VRAM. For models up to 3B parameters, we carry out both fine-tuning and few-shot prompting on a single GPU. For the 8B model, we use two GPUs for few-shot prompting, and for the 70B model, we rely on eight GPUs.

When fine-tuning the RoBERTa model, we set the learning rate to 2×10^{-5}, use a batch size of 8, and train for 10 epochs. For the LLaMA 3.2 3B model, we set the learning rate to 2×10^{-5}, use a batch size of 4, and train for 5 epochs. During few-shot prompting with LLMs, we use the default maximum model input length without imposing additional constraints.