On Distributional Robustness of In-Context Learning for Text Classification

Anonymous ACL submission

Abstract

Pretrained language models (LMs) have been shown to be capable of few-shot learning via in-context demonstrations. In this work, we examine if in-context learning generalizes out-of-distribution. On several text classification tasks and considering different kinds of realistic demonstration vs. test distribution shifts—domain, adversarial, and dialectal—we find a significant drop in test accuracy compared to in-distribution performance. Moreover, we find that the accuracy is inversely proportional to the number of demonstrations. To address this issue, we explore different zero-shot prompting approaches which do not rely on demonstrations. With six open-source language model families, we show that zero-shot prompting techniques which verbalize the target distribution in the prompt are able to close the gap to in-domain few-shot classification performance.

1 Introduction

Language models trained on only raw text have been shown to learn in-context, that is, perform new tasks simply by conditioning on a handful of demonstrations (Brown et al., 2020). By drawing parallels to gradient-based optimization (Dai et al., 2023; Deutch et al., 2023), prior work has suggested that in-context learning (ICL) is prone to relying on superficial features in the training examples (Mueller et al., 2023b). Hence, it can be brittle to distribution shifts between the demonstrations and test examples with more demonstrations and larger models showing sharper drops in task performance (Tang et al., 2023).

With text classification as a case study, in this work, we investigate the utility of zero-shot inference methods in mitigating that issue, as they do not rely on any demonstrations. Recent studies have suggested that ICL serves as a way to implicitly prime the model with the domain, concepts, or topics and the format of the target task (Min et al., 2022; Wang et al., 2023). We build on the approach presented in Kumar et al. (2024) which showed that specifying these concepts explicitly in the prompt, without any demonstrations, can match or even surpass few-shot performance.

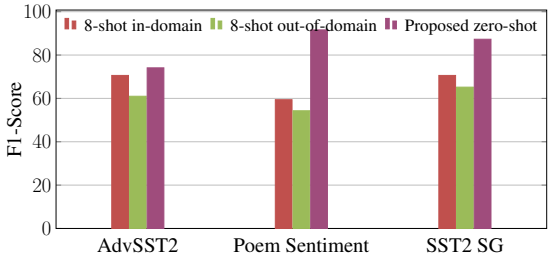

Figure 1: We observe considerable performance (F1) drops for sentiment classification when testing with three kinds of distribution shifts: adversarial, domain (poetry), and dialect (Singlish) when tested with Mistral 7B. The demonstrations are from SST2. Our proposed approach (§4) is able to close the gap.

Evaluating on text classification tasks with three kinds of realistic distribution shifts (adversarial, domain, and dialectal) and with six language model families (model sizes ranging from 125M to 13B), we find that simply verbalizing the distributional properties of the test examples in the prompt, such as its domain or dialect, results in zero-shot approaches matching 8-shot in-domain performance. In a more realistic setting where the test distribution may not be known, we propose a simple but effective *mixture of prompts* approach that computes and aggregates model probabilities based on prompts containing verbalizations of different distributional shifts. Our experiments show that when the prompt mixture contains the true distribution information of the test example, this approach performs on par with the case where the properties are known in advance. In cases where the prompt mixture does not contain the test distribution description, the performance declines, however it still outperforms few-shot approaches.

	0 shot		1 shot		4 shot		8 shot		16 shot	
	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD
Poem	$54.3_{0.0}$	$54.3_{0.0}$	$85.7_{4.8}$	$85.7_{13.3}$	$92.6_{1.4}$	$80.0_{9.9}$	$93.1_{4.6}$	$52.6_{3.4}$	$92.6_{3.9}$	$88.0_{2.8}$
Finance	$92.3_{0.0}$	$92.3_{0.0}$	$94.1_{3.7}$	$95.1_{0.6}$	$96.2_{0.5}$	$90.4_{5.0}$	$95.7_{1.4}$	$74.3_{32.5}$	$93.3_{1.4}$	$93.3_{5.1}$

Table 1: Few-shot Macro-F1 (mean_{std}) with Pythia 6.9B for domain shifts with demonstrations from SST2. In-Distribution (ID) refers to cases where test cases and training data where from the same domain, and Out-of-Distribution (OOD) refers to cases where they were from different domains (see Appendix B for additional results).

	0 shot			1 shot			4 shot			8 shot		
	SAE	AAVE	SG	SAE	AAVE	SG	SAE	AAVE	SG	SAE	AAVE	SG
SST2	$50.7_{0.0}$	$52.9_{0.0}$	$52.5_{0.0}$	$88.9_{3.2}$	$88.9_{2.4}$	$86.4_{2.6}$	$93.6_{0.8}$	$90.8_{2.1}$	$89.3_{1.8}$	$93.6_{1.3}$	$90.4_{2.8}$	$88.9_{3.3}$
Emotion	$42.0_{0.0}$	$39.6_{0.0}$	$38.7_{0.0}$	$34.6_{3.4}$	$33.5_{4.2}$	$31.6_{4.9}$	$31.8_{7.5}$	$30.3_{8.0}$	$29.6_{8.8}$	$33.2_{5.3}$	$32.3_{5.8}$	$31.4_{6.1}$

Table 2: Few-shot Macro-F1 (mean_{std}) with Pythia 6.9B for dialect shifts from Standard American (SAE) to African American Vernacular (AAVE) and Singaporean English dialects (SG; see Appendix B for additional results).

2 Related Work

Since its introduction, various studies have attempted to analyze ICL's underlying mechanisms (Xie et al., 2022; Ahuja et al., 2023; Hahn and Goyal, 2023; Zhang et al., 2023; von Oswald et al., 2023; Wang et al., 2023) as well as explored its limits (Akyürek et al., 2022; Chan et al., 2022). Prior work has shown that ICL relies on shortcuts to make predictions (Tang et al., 2023). Saparov and He (2023) find that LMs are tends to rely on spurious correlations for reasoning tasks that hinder robust generalization. However, they have largely focused on toy or synthetic tasks. In this work, we focus on realistic distribution shifts on popular text classification tasks such as sentiment, topic and emotion widely studied in the context of ICL. Most related to our work is Mueller et al. (2023a), which tests generalization for syntactic tasks.

While most prior work only points out the issue of robustness of ICL, we also explore potential solutions to tackle this issue using zero-shot approaches which in recent work have shown to match few-shot performance for text classification tasks Kumar et al. (2024). Relatedly, Drozdov et al. (2023), find that a series of chain-of-thought prompts can yield better robustness for semantic parsing.

3 ICL Is Not Distributionally Robust

In an in-context learning setup, given k demonstrations $\{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_k, y_k)\}$ and a test example $\mathbf{x}_{\text{test}}$, the label is predicted using a language model,

$$\max_{y \in \mathcal{Y}} p_{\text{LM}}(y \mid \mathbf{x}_1, y_1, \ldots, \mathbf{x}_k, y_k, \mathbf{x}_{\text{test}}) \quad (1)$$

Here $\mathcal{Y}$ is the set of all labels. In practice, the labels are verbalized in natural language (e.g., the words "negative" and "positive" for sentiment classification) and all the demonstrations are concatenated followed by the test example. Typically, all $\mathbf{x}_i$ and $\mathbf{x}_{\text{test}} \sim \mathcal{D}$, an input distribution. Text classifiers *trained* with supervision have been shown to learn shortcuts in the data to make predictions which limit their generalization. In this section, we examine this phenomenon for ICL by measuring if the classification accuracy is impacted when the test distribution diverges from the demonstrations, that is when $\mathbf{x}_{\text{test}} \sim \mathcal{D}'$ different from $\mathcal{D}$.

Experimental Setup and Results We evaluate on three tasks: sentiment (2-class), emotion (6-class), and topic (4-class) classification with demonstrations from SST2 (movie reviews; Socher et al., 2013), Emotions (Saravia et al., 2018), and AGNews (Zhang et al., 2015) respectively. We consider two kinds of distribution shifts in test examples: domain and dialectal. For domain shifts, we evaluate on two binary sentiment classification datasets: Poetry (Sheng and Uthus, 2020), and Financial News (Malo et al., 2014). We simulate the out-of-distribution (OOD) setup using demonstrations from the SST2 train set and compare with in-domain demonstrations. For dialectal shifts, we use multi-VALUE toolkit (Ziems et al., 2023) to translate SST2, Emotions, and AGNews to African American Vernacular English (AAVE) and Singaporean English (SG) using the original train set examples as demonstrations. We report macro F1-scores for all evaluation sets for $k \in \{0, 1, 4, 8\}$ where for each k we report mean and variance with 10 different sets of demonstrations sampled from the respective train sets. We evaluated on GPT2 (S, M, L, XL) (Radford et al., 2019), OPT (1.3B and

2.7B) (Zhang et al., 2022), Pythia (1.4B, 2.8B, and 6.9B) (Biderman et al., 2023), Mistral 7B (Jiang et al., 2023), Llama 1 (7B) (Touvron et al., 2023), and Llama 2 (7B and 13B) (Touvron et al., 2023). No finetuning was performed on these models.

We report results for Pythia 6.9B in Tables 1 and 2 (with results for other models following largely similar trends in Appendix B). We observe a significant decline in OOD accuracy for both domain and dialectal shifts with the former being much larger. Furthermore, while more demonstrations increase in-domain performance, we observe an inverse trend with OOD setups.

4 Context Aware Zero-Shot Inference

As shown in Tables 1 and 2, the zero-shot setup observes no decline in performance across distributions since, by definition, it does not rely on any demonstrations. However, its overall accuracy is poor even in-domain compared to higher-shot setups. To close this gap, in this section, we describe *context-aware* zero-shot inference which uses the following inference objective: $\hat{y} = \max_{y \in \mathcal{Y}} p(y \mid \mathbf{x}_{\text{test}}, u, v, \ldots)$. Here, $u, v, \ldots$ represent additional context that can aid the task (such as u could be domain of the input text, v could be the author demographic, and so on). Kumar et al. (2024), who proposed this approach, experimented with different instantiations of these factors showing that conditioning on the domain, subject, author, or reader of the input text can help improve classification performance to match up to 16-shot results while achieving low variance across prompt variations.

In this work, we explore the utility this setup where we consider the type of test distribution as additional context and verbalize it along with the label. We use a generative version of this setup, where $\hat{y} = \max_{y \in \mathcal{Y}} p(\mathbf{x}_{\text{test}} \mid y, u, v, \ldots)$ to make predictions, which has been shown to perform better than the previously described discriminative setup. We verbalize the labels and the contextual factors $(y, u, v, \ldots)$ in textual form and refer to it as z (for example, "this is a positive^{y} movie review^{u} written in a Singaporean English dialect^{v}"). Furthermore, to reduce variance across prompt variations, we aggregate model probabilities across prompt paraphrases (Kumar et al., 2024) to obtain:

$$\hat{y}_{\text{GEN}} = \max_{y \in \mathcal{Y}} \sum_{z \in \mathcal{Z}(y,u,v,\ldots)} p(\mathbf{x} \mid z) \qquad (2)$$

$\mathcal{Z}$ describes all potential ways to verbalize the label and the factors (in practice, we only use 10). However, the factors, i.e., the test distribution, might not be known. To overcome this issue, we propose to use a *mixture of prompts* which aggregates the probabilities in (2) over all *potential* factors,

$$\hat{y}_{\text{GEN}} = \max_{y \in \mathcal{Y}} \sum_{u,v,\ldots} \sum_{z \in \mathcal{Z}(y,u,v,\ldots)} p(\mathbf{x}|z) \qquad (3)$$

The set of enumerable distributions can be infinitely large. In our experiments, we simulate a setting with a small set where we compare performance when the target distribution is either included or excluded from the set.

4.1 Experimental Setup

In addition to all the datasets described in §3, we evaluate on additional domains for sentiment classification, namely, CR (Customer Reviews; Hu and Liu, 2004), and MR (RottenTomatoes; Pang and Lee, 2005). We also experiment with an adversarial distribution shift by evaluating on adv-SST2 (Wang et al., 2022) (more details in Appendix A).

We consider three scenarios. The first is where we know the distribution of the test set (**known**). In this scenario, we make predictions using (2). In the second scenario, we do not know the distribution of the test set, but we assume that it exists in the set of potential distributions (**partly known**). In the third scenario, we have a completely unknown distribution that is not present in the set (**unknown**). For both of these scenarios, we use (3) for inference.

For each scenario, task and dataset, we hand write a label description z per label per distribution. We then generate 10 templatic paraphrases by querying ChatGPT (GPT-3.5) (Ouyang et al., 2022)[1] and manually verify the correctness of each paraphrase. For *mixture of prompt* experiments, for domain and adversarial shifts, we include the following distributions in the mixture: movie reviews, Amazon reviews, Yelp reviews, Poetry, Financial news, adversarial reviews. For dialectal shifts for sentiment, emotion and topic, we evaluate on AAVE and Singaporean English versions of the test sets and include AAVE, Singaporean, Indian, Irish, and Nigerian English in the mixture. Note that we are only including the names of these distributions in the prompts, and not any demonstrations. Finally, we simulate a case where we do

[1]This process needs to be done only once for each task and, in practice, any paraphrasing model can be employed. We provide the list of all paraphrases we generated here: https://pastebin.com/2gBYYxJU

	Few-shot(OOD)	**Zero-shot**(Known)	**Zero-shot**(Partly Known)	**Zero-shot**(Unknown)
Adv-SST2	$62.7_{5.8}$	$62.2_{2.2}$	$73.9_{0.6}$	$\mathbf{74.1}_{0.8}$
MR	$70.2_{10.7}$	$\mathbf{90.2}_{0.2}$	$86.4_{0.2}$	$81.3_{0.2}$
CR	$71.5_{8.1}$	$86.1_{0.3}$	$\mathbf{87.8}_{0.4}$	$76.4_{0.6}$
Amazon	$82.9_{5.1}$	$92.3_{0.3}$	$\mathbf{93.0}_{0.2}$	$92.3_{0.1}$
Yelp	$82.5_{6.8}$	$\mathbf{93.7}_{0.1}$	$92.7_{0.1}$	$92.8_{0.1}$
Poem	$59.4_{6.9}$	$\mathbf{91.6}_{0.8}$	$89.6_{0.8}$	$90.0_{0.0}$
Finance	$90.6_{0.0}$	$\mathbf{97.1}_{0.6}$	$91.3_{0.9}$	$90.7_{0.6}$
SST2 AAVE	$71.4_{10.1}$	$88.8_{0.3}$	$89.1_{0.1}$	$\mathbf{89.3}_{0.2}$
SST2 SG	$69.3_{9.1}$	$87.3_{0.2}$	$\mathbf{88.6}_{0.2}$	$88.6_{0.3}$
Emotion AAVE	$40.9_{2.6}$	$\mathbf{51.4}_{0.2}$	$49.6_{0.2}$	$48.0_{0.1}$
Emotions SG	$40.8_{2.5}$	$48.0_{0.2}$	$46.5_{0.1}$	$\mathbf{49.2}_{0.1}$
AGNews AAVE	$\mathbf{82.6}_{0.0}$	$48.1_{0.1}$	$48.1_{0.0}$	$47.5_{0.0}$
AGNews SG	$\mathbf{89.9}_{8.2}$	$48.3_{0.1}$	$47.6_{0.0}$	$47.4_{0.0}$

Table 3: Macro-F1 with Mistral 7B. We report mean_{std} over 10 runs for zero-shot approaches over 5 seeds for the few-shot ones. For each result the best approach within that category is shown; Detailed results in Appendix B.

not know the test distribution, by removing from the mixture, the prompts related to the test distribution. For each test set, we report mean and standard deviation of macro-F1.

In addition to the generative classification approach in (3), we also evaluate on the discriminative version defined by $p(z \mid \mathbf{x})$. We consider three versions of this approach which differ in how the context is provided (details in Appendix A)

For few-shot baselines, we consider two versions: one which uses (1) (ICL), and another which calibrates the label probability by dividing it with $p_{\text{LM}}(y \mid \text{NULL})$ (ICL-CC) (Zhao et al., 2021).

4.2 Results

We report representative results in Table 3 for Mistral 7B with the best results for each of the three zero-shot scenarios and the best performing few-shot baseline (with detailed breakdown for all models with largely similar trends in Appendix B).

We find that if the test distribution is known, the zero-shot approaches achieve the best performance overall surpassing the few-shot out-of-distribution accuracy by a large margin and even approaching the in-distribution test performance (see Figure 1). This trend holds for the majority of the 12 models we tested (in 8-10 cases depending on the dataset).

In a more realistic scenario, if the test distribution is unknown, the mixture of prompts approach described in (3) (partially unknown) still performs better than the few-shot baselines. However, the improvements are not as large compared to the first scenario as explicitly priming the model with a different distribution than the target might hurt. When the target distribution is not included in the prompt mixture (unknown), we find that this approach still performs better than baseline, however, the accuracy gain is lower than previous two scenarios. While "unknown" depicts the worst case scenario, for most practical purposes, given a large enough of set of distributions, the target distribution is highly likely to be included in the set.

Finally, the generative zero-shot classification objective is the winning approach in most settings.[2] However, the discriminative versions are competitive and even out-perform the generative approach in a few cases (see details in Appendix A).

5 Conclusion

In this work, we investigate the out-of-distribution robustness of in-context learning. On several text classification tasks we find that a distribution mismatch between the test instances and the demonstrations can lead to a considerable drop in accuracy compared to in-distribution accuracy. We propose to address this issue using zero-shot prompting. Our experiments show that simply describing the distribution in the prompt text can considerably close the performance gap.

[2]The AGNews dataset does not reflect these results. This is likely because our dialectal data creation method does not modify topical words in the test examples.

6 Limitations

The datasets in our study are exclusively in English, so the generalizability of our paper's findings to other languages may be limited. In addition, our dialectical distribution datasets are synthetically generated, so they may not perfectly reflect the dialects they represent. We also make simplifying assumptions about label independence which may not be true in practice. Due to computing limitations, we were unable to run experiments on larger models, which may not indicate the same trends.

References

Kabir Ahuja, Madhur Panwar, and Navin Goyal. 2023. In-context learning through the bayesian prism. *Preprint*, arXiv:2306.04891.

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

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Stephanie CY Chan, Ishita Dasgupta, Junkyung Kim, Dharshan Kumaran, Andrew K Lampinen, and Felix Hill. 2022. Transformers generalize differently from information stored in context vs in weights. *arXiv preprint arXiv:2210.05675*.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. 2023. Why can GPT learn in-context? language models secretly perform gradient descent as meta-optimizers. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4005–4019, Toronto, Canada. Association for Computational Linguistics.

Gilad Deutch, Nadav Magar, Tomer Bar Natan, and Guy Dar. 2023. In-context learning and gradient descent revisited. *arXiv e-prints*, pages arXiv–2311.

Andrew Drozdov, Nathanael Schärli, Ekin Akyürek, Nathan Scales, Xinying Song, Xinyun Chen, Olivier Bousquet, and Denny Zhou. 2023. Compositional semantic parsing with large language models. In *The Eleventh International Conference on Learning Representations*.

Michael Hahn and Navin Goyal. 2023. A theory of emergent in-context learning as implicit structure induction. *Preprint*, arXiv:2303.07971.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Sachin Kumar, Chan Young Park, and Yulia Tsvetkov. 2024. Gen-z: Generative zero-shot text classification with contextualized label descriptions. In *The Twelfth International Conference on Learning Representations*.

Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796.

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? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Aaron Mueller, Kanika Narang, Lambert Mathias, Qifan Wang, and Hamed Firooz. 2023a. Meta-training with demonstration retrieval for efficient few-shot learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6049–6064, Toronto, Canada. Association for Computational Linguistics.

Aaron Mueller, Albert Webson, Jackson Petty, and Tal Linzen. 2023b. In-context learning generalizes, but not always robustly: The case of syntax. *ArXiv*, abs/2311.07811.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *arXiv preprint cs/0506075*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Abulhair Saparov and He He. 2023. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. *Preprint*, arXiv:2210.01240.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. Carer: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3687–3697.

Emily Sheng and David Uthus. 2020. Investigating societal biases in a poetry composition system. *arXiv preprint arXiv:2011.02686*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Ruixiang Tang, Dehan Kong, Longtao Huang, and Hui Xue. 2023. Large language models can be lazy learners: Analyze shortcuts in in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4645–4657, Toronto, Canada. 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*.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. 2023. Transformers learn in-context by gradient descent. *Preprint*, arXiv:2212.07677.

Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. 2022. Adversarial glue: A multi-task benchmark for robustness evaluation of language models. *Preprint*, arXiv:2111.02840.

Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. 2023. Large language models are implicitly topic models: Explaining and finding good demonstrations for in-context learning. In *Workshop on Efficient Systems for Foundation Models@ ICML2023*.

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*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Yufeng Zhang, Fengzhuo Zhang, Zhuoran Yang, and Zhaoran Wang. 2023. What and how does in-context learning learn? bayesian model averaging, parameterization, and generalization. *Preprint*, arXiv:2305.19420.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. PMLR.

Caleb Ziems, William Held, Jingfeng Yang, Jwala Dhamala, Rahul Gupta, and Diyi Yang. 2023. Multi-VALUE: A framework for cross-dialectal English NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 744–768, Toronto, Canada. Association for Computational Linguistics.

A Additional Experimental Details

Discriminative Baselines:

We consider three variations of each discriminative zero-shot baseline (see Table 18): (1) Direct context specifies the contextual variable using a simple format, (2) Direct instruct specifies the contextual variable and also provides the labels the model is expected to predict, and (3) Direct simple does not provide the contextual variable.

Table 4 summarizes all of the datasets we use. Table 17 summarizes the handwritten label description templates we use and paraphrase to construct the full set of label descriptions.

B Detailed Results

Detailed results for the following models are provided in this Appendix: GPT2 Results in Table 5, GPT2-Medium Results in Table 6, GPT2-Large Results in Table 7, GPT2-XL Results in Table 8, OPT 1.3B Results in Table 9, Pythia 1.4B Results in Table 10, OPT 2.7B Results in Table 11, Pythia 2.8B Results in Table 12, Pythia 6.9B Results in Table 13, Mistral 7B Results in Table 14, Llama 7B Results in Table 15, and Llama-2 7B Results in Table 16.

Dataset	Task	Domain	# of Classes	Inclusion in Minimal Set
SST2	Sentiment Classification	Movies	2	Yes
ADV SST2	Sentiment Classification	Movies, adversarial	2	Yes
MR	Sentiment Classification	Movies, Rotten Tomatoes	2	No
CR	Sentiment Classification	Customer Reviews	2	Yes
Amazon	Sentiment Classification	Customer Reviews, Amazon	2	No
Yelp	Sentiment Classification	Yelp reviews	2	Yes
Poem	Sentiment Classification	Poetry	4, reduced to 2	Yes
Finance	Sentiment Classification	Economic News	3, reduced to 2	Yes
SST2 AAVE	Sentiment Classification	Movies, African American Vernacular English (AAVE), synthetically created	2	Yes
SST2 SG	Sentiment Classification	movies, Singaporean English (Singlish), synthetically created)	2	Yes
Emotion AAVE	Emotion Classification	Tweets, African American Vernacular English (AAVE), synthetically created	6	Yes
Emotion SG	Emotion Classification	Tweets, Singaporean English (Singlish), synthetically created)	6	Yes
AG News AAVE	Topic Classification	News, African American Vernacular English (AAVE), synthetically created	4	Yes
AG News SG	Topic Classification	News, Singaporean English (Singlish), synthetically created)	4	Yes

Table 4: Datasets used for the experiments

	Channel simple, ID	Direct context, ID	Direct instruct, ID	Direct simple, ID	Channel simple, ID mix	Direct context, ID mix	Direct instruct, ID mix	Direct simple, ID mix	Channel simple, OOD mix	Direct context, OOD mix	Direct instruct, OOD mix	Direct simple, OOD mix	0-shot, OOD	1-shot, OOD	4-shot, OOD	8-shot, OOD
SST2	$83.1_{0.3}$	$72.8_{0.5}$	$58.4_{0.3}$	$74.5_{0.6}$	$68.3_{0.3}$	$73.9_{0.2}$	$77.63_{0.2}$	$73.6_{0.5}$	$64.5_{0.4}$	$72.7_{0.4}$	$61.0_{0.4}$	$70.2_{0.2}$	$66.5_{0.0}$	$57.5_{3.3}$	$58.6_{7.8}$	$55.9_{6.1}$
SST2 ADV	$42.1_{1.7}$	$46.6_{1.1}$	$43.3_{2.1}$	$44.7_{1.4}$	$48.1_{0.7}$	$50.9_{0.8}$	$43.1_{0.4}$	$43.6_{0.6}$	$48.2_{0.6}$	$49.5_{1.4}$	$44.6_{0.6}$	$42.7_{0.8}$	$44.6_{0.0}$	$47.4_{3.3}$	$49.2_{1.7}$	$46.9_{2.3}$
MR	$76.9_{0.3}$	$66.2_{0.6}$	$53.8_{0.5}$	$64.8_{0.6}$	$66.3_{0.5}$	$70.0_{0.2}$	$72.6_{0.2}$	$70.1_{0.4}$	$63.1_{0.3}$	$66.9_{0.2}$	$58.8_{0.4}$	$65.8_{0.4}$	$62.4_{0.0}$	$55.5_{4.0}$	$53.1_{2.7}$	$55.7_{6.8}$
CR	$73.4_{0.3}$	$69.2_{0.7}$	$40.0_{0.4}$	$55.1_{0.9}$	$63.6_{0.7}$	$61.9_{0.4}$	$45.1_{0.2}$	$69.4_{0.7}$	$63.2_{0.8}$	$63.0_{0.7}$	$45.0_{0.4}$	$71.1_{0.6}$	$73.4_{0.0}$	$54.0_{6.0}$	$40.4_{4.3}$	$45.1_{11.8}$
Amazon	$78.1_{0.4}$	$77.0_{0.6}$	$62.4_{0.6}$	$73.4_{0.6}$	$76.5_{0.7}$	$74.5_{0.3}$	$84.1_{0.1}$	$76.2_{0.3}$	$76.6_{0.4}$	$75.2_{0.3}$	$84.0_{0.1}$	$77.3_{0.3}$	$67.1_{0.0}$	$56.5_{2.3}$	$57.9_{3.4}$	$61.4_{6.6}$
Yelp	$80.4_{0.3}$	$82.5_{0.5}$	$62.8_{0.3}$	$82.2_{0.4}$	$77.7_{0.2}$	$77.6_{0.1}$	$82.5_{0.2}$	$76.9_{0.2}$	$78.2_{0.3}$	$76.8_{0.2}$	$82.6_{0.2}$	$77.7_{0.2}$	$75.1_{0.0}$	$54.3_{4.4}$	$57.2_{3.0}$	$62.8_{9.8}$
Poem	$85.6_{2.3}$	$52.8_{3.7}$	$48.0_{0.0}$	$56.2_{4.3}$	$78.2_{2.2}$	$74.6_{1.0}$	$58.0_{0.0}$	$68.4_{0.8}$	$79.4_{3.0}$	$75.8_{2.2}$	$58.0_{0.0}$	$68.0_{0.0}$	$77.1_{0.0}$	$51.4_{6.0}$	$57.7_{4.2}$	$54.3_{0.0}$
Finance	$79.8_{1.0}$	$75.1_{1.6}$	$47.5_{2.3}$	$72.0_{1.2}$	$71.4_{0.7}$	$59.5_{0.9}$	$40.9_{0.5}$	$65.3_{1.2}$	$70.8_{0.6}$	$58.9_{0.7}$	$41.3_{0.4}$	$65.0_{0.4}$	$50.2_{0.0}$	$29.6_{7.0}$	$47.7_{24.8}$	$90.6_{0.0}$
SST2 AAVE	$80.3_{0.3}$	$54.5_{0.4}$	$60.5_{0.6}$	$65.9_{0.6}$	$80.4_{0.3}$	$50.9_{0.1}$	$50.9_{0.1}$	$51.0_{0.1}$	$80.1_{0.2}$	$58.4_{0.6}$	$59.4_{1.0}$	$67.5_{0.3}$	$64.0_{0.0}$	$55.1_{4.3}$	$52.2_{0.9}$	$49.1_{0.0}$
SST2 SG	$78.2_{0.4}$	$59.6_{0.6}$	$51.5_{0.2}$	$67.2_{0.6}$	$77.9_{0.2}$	$50.9_{0.1}$	$50.9_{0.1}$	$50.9_{0.1}$	$78.1_{0.2}$	$56.9_{0.3}$	$62.8_{0.3}$	$65.7_{0.4}$	$64.1_{0.0}$	$56.1_{3.5}$	$51.6_{2.0}$	$53.9_{6.0}$

Table 5: Macro-F1 with GPT2. We report $\mathrm{mean}_{\mathrm{std}}$ over 10 runs for zero-shot. For few-shot, we report $average_{std}$ over 5 seeds.

	Channel simple, ID	Direct context, ID	Direct instruct, ID	Direct simple, ID	Channel simple, ID mix	Direct context, ID mix	Direct instruct, ID mix	Direct simple, ID mix	Channel simple, OOD mix	Direct context, OOD mix	Direct instruct, OOD mix	Direct simple, OOD mix	0-shot, OOD	1-shot, OOD	4-shot, OOD	8-shot, OOD
SST2	$88.6_{3.2}$	$76.6_{0.4}$	$81.5_{0.5}$	$73.1_{0.5}$	$80.1_{0.3}$	$77.3_{0.1}$	$80.0_{0.4}$	$77.6_{0.3}$	$76.6_{0.5}$	$83.3_{0.2}$	$75.3_{0.2}$	$81.3_{0.3}$	$77.0_{0.0}$	$60.1_{9.9}$	$58.3_{13.3}$	$63.2_{12.3}$
SST2 ADV	$44.2_{1.5}$	$44.0_{1.0}$	$48.9_{1.4}$	$42.2_{2.1}$	$51.5_{1.2}$	$47.0_{0.9}$	$51.3_{0.8}$	$44.8_{0.8}$	$50.6_{0.8}$	$45.7_{0.6}$	$50.1_{0.5}$	$44.2_{0.8}$	$43.9_{0.0}$	$45.4_{3.4}$	$48.0_{2.1}$	$48.8_{3.2}$
MR	$80.1_{0.30}$	$70.0_{1.0}$	$79.2_{0.6}$	$66.5_{0.8}$	$75.0_{0.3}$	$73.8_{0.3}$	$77.1_{0.2}$	$77.8_{0.2}$	$71.4_{0.3}$	$80.1_{0.3}$	$73.7_{0.3}$	$78.4_{0.2}$	$73.9_{0.0}$	$58.7_{8.7}$	$57.6_{11.9}$	$63.2_{10.4}$
CR	$84.0_{0.3}$	$85.1_{1.0}$	$51.4_{1.0}$	$73.2_{1.3}$	$78.3_{0.6}$	$61.6_{0.9}$	$45.2_{0.2}$	$63.9_{0.4}$	$78.8_{0.6}$	$61.2_{1.0}$	$45.0_{0.3}$	$64.0_{0.8}$	$78.2_{0.0}$	$62.3_{11.4}$	$52.9_{17.3}$	$74.3_{6.7}$
Amazon	$62.0_{0.3}$	$86.7_{0.4}$	$75.3_{0.5}$	$83.2_{0.4}$	$75.1_{0.3}$	$80.5_{0.3}$	$80.9_{0.1}$	$82.1_{0.3}$	$76.3_{0.4}$	$80.8_{0.2}$	$80.7_{0.1}$	$82.4_{0.3}$	$75.7_{0.0}$	$61.1_{9.1}$	$65.7_{8.9}$	$66.8_{16.1}$
Yelp	$85.8_{0.2}$	$90.3_{0.1}$	$73.2_{0.4}$	$89.1_{0.2}$	$80.7_{0.2}$	$77.6_{0.2}$	$81.3_{0.1}$	$77.2_{0.2}$	$80.3_{0.3}$	$77.5_{0.2}$	$81.2_{0.1}$	$77.1_{0.2}$	$87.9_{0.0}$	$65.4_{11.0}$	$70.6_{10.9}$	$67.2_{17.8}$
Poem	$84.8_{1.7}$	$63.0_{2.2}$	$48.2_{0.6}$	$77.6_{2.5}$	$71.4_{2.1}$	$77.4_{1.0}$	$80.0_{0.0}$	$76.0_{0.9}$	$72.0_{1.9}$	$75.8_{0.6}$	$80.0_{0.0}$	$76.0_{1.3}$	$77.1_{0.0}$	$59.4_{7.1}$	$54.3_{6.5}$	$54.3_{0.0}$
Finance	$90.0_{0.5}$	$90.7_{1.4}$	$59.9_{2.4}$	$86.1_{1.0}$	$82.7_{0.4}$	$77.5_{0.6}$	$84.2_{0.3}$	$71.0_{0.9}$	$82.4_{0.9}$	$76.8_{0.9}$	$85.0_{0.9}$	$73.8_{0.8}$	$86.5_{0.0}$	$74.6_{32.6}$	$68.2_{31.5}$	$90.6_{0.0}$
SST2 AAVE	$83.4_{0.5}$	$71.6_{0.6}$	$72.4_{0.4}$	$75.7_{0.8}$	$83.8_{0.2}$	$49.1_{0.0}$	$49.1_{0.0}$	$49.1_{0.0}$	$82.4_{0.2}$	$75.2_{0.3}$	$53.7_{0.2}$	$73.5_{0.2}$	$75.1_{0.0}$	$60.1_{9.0}$	$57.5_{12.9}$	$63.4_{10.9}$
SST2 SG	$79.0_{0.3}$	$64.1_{0.7}$	$59.2_{0.4}$	$73.7_{0.4}$	$81.0_{0.2}$	$49.1_{0.0}$	$49.1_{0.0}$	$49.1_{0.0}$	$81.4_{0.1}$	$70.6_{0.4}$	$59.0_{0.4}$	$69.9_{0.4}$	$74.8_{0.0}$	$61.0_{8.9}$	$56.4_{12.0}$	$62.0_{9.6}$

Table 6: Macro-F1 with GPT2-Medium. We report $average_{std}$ over 10 runs for zero-shot. For few-shot, we report $average_{std}$ over 5 seeds.

	Channel simple, ID	Direct context, ID	Direct instruct, ID	Direct simple, ID	Channel simple, ID mix	Direct context, ID mix	Direct instruct, ID mix	Direct simple, ID mix	Channel simple, OOD mix	Direct context, OOD mix	Direct instruct, OOD mix	Direct simple, OOD mix	0-shot, OOD	1-shot, OOD	4-shot, OOD	8-shot, OOD
SST2	$88.6_{0.2}$	$84.3_{0.5}$	$79.9_{0.3}$	$83.5_{0.4}$	$81.6_{0.3}$	$83.8_{0.4}$	$76.6_{0.1}$	$83.1_{0.2}$	$75.2_{0.3}$	$83.6_{0.1}$	$75.0_{0.3}$	$83.6_{0.2}$	$69.8_{0.0}$	$74.1_{11.5}$	$73.7_{10.1}$	$81.0_{9.5}$
SST2 ADV	$47.6_{0.9}$	$51.1_{1.2}$	$58.9_{1.1}$	$51.2_{1.8}$	$52.3_{1.0}$	$47.8_{1.0}$	$51.0_{0.4}$	$51.1_{1.0}$	$52.9_{0.9}$	$47.5_{0.8}$	$51.6_{0.9}$	$53.4_{0.7}$	$43.2_{0.0}$	$41.5_{5.4}$	$52.6_{2.5}$	$49.0_{3.3}$
MR	$82.9_{0.2}$	$72.4_{0.6}$	$78.0_{0.5}$	$72.2_{0.5}$	$76.5_{0.3}$	$78.2_{0.3}$	$75.6_{0.1}$	$78.6_{0.2}$	$71.8_{0.5}$	$80.6_{0.2}$	$71.5_{0.5}$	$79.2_{0.3}$	$76.2_{0.0}$	$72.5_{11.1}$	$67.5_{9.2}$	$80.3_{10.7}$
CR	$82.6_{0.4}$	$83.4_{0.7}$	$74.1_{1.1}$	$80.4_{0.5}$	$73.9_{0.8}$	$77.9_{0.5}$	$47.7_{0.2}$	$79.0_{0.8}$	$72.9_{0.7}$	$72.8_{0.4}$	$47.6_{0.2}$	$76.2_{0.8}$	$84.8_{0.0}$	$76.3_{6.4}$	$71.7_{4.6}$	$78.3_{6.7}$
Amazon	$84.6_{0.4}$	$86.5_{0.4}$	$83.1_{0.5}$	$80.0_{0.5}$	$85.8_{0.4}$	$86.9_{0.2}$	$74.7_{0.1}$	$87.9_{0.3}$	$85.8_{0.4}$	$86.6_{0.2}$	$74.7_{0.1}$	$87.3_{0.3}$	$90.6_{0.0}$	$80.9_{13.5}$	$79.8_{3.8}$	$81.5_{5.3}$
Yelp	$87.1_{0.2}$	$90.5_{0.3}$	$80.2_{0.4}$	$89.1_{0.4}$	$85.3_{0.3}$	$75.9_{0.2}$	$70.9_{0.2}$	$76.0_{0.3}$	$85.0_{0.3}$	$74.6_{0.2}$	$70.9_{0.1}$	$74.9_{0.3}$	$94.7_{0.0}$	$80.3_{15.2}$	$76.3_{3.2}$	$74.6_{7.1}$
Poem	$86.6_{1.6}$	$80.6_{2.5}$	$46.4_{0.8}$	$84.6_{3.3}$	$70.6_{2.3}$	$84.0_{1.6}$	$52.0_{0.9}$	$77.8_{1.1}$	$73.2_{1.7}$	$84.6_{1.3}$	$51.6_{1.3}$	$77.8_{1.1}$	$80.0_{0.0}$	$56.0_{5.9}$	$45.7_{4.8}$	$52.6_{3.4}$
Finance	$85.7_{0.9}$	$92.6_{0.9}$	$37.4_{0.9}$	$83.3_{1.1}$	$73.2_{0.6}$	$85.6_{0.4}$	$47.7_{0.6}$	$82.5_{0.8}$	$73.1_{0.4}$	$87.8_{0.6}$	$47.6_{0.6}$	$86.2_{0.4}$	$92.3_{0.0}$	$70.4_{31.7}$	$61.1_{30.1}$	$74.3_{32.5}$
SST2 AAVE	$83.8_{0.3}$	$52.5_{0.2}$	$50.7_{0.2}$	$82.3_{0.8}$	$83.5_{0.1}$	$49.1_{0.0}$	$49.1_{0.0}$	$49.1_{0.0}$	$84.3_{0.3}$	$78.1_{0.3}$	$57.1_{0.3}$	$80.9_{0.4}$	$76.2_{0.0}$	$72.6_{10.7}$	$67.7_{10.0}$	$79.8_{11.7}$
SST2 SG	$81.7_{0.2}$	$55.9_{0.4}$	$51.3_{0.2}$	$78.4_{0.4}$	$84.1_{0.2}$	$49.1_{0.0}$	$49.1_{0.0}$	$49.1_{0.0}$	$83.9_{0.1}$	$76.5_{0.6}$	$53.2_{0.1}$	$80.6_{0.4}$	$73.3_{0.0}$	$73.3_{11.1}$	$66.3_{10.3}$	$79.2_{11.9}$

Table 7: Macro-F1 with GPT2-Large. We report $average_{std}$ over 10 runs for zero-shot. For few-shot, we report $average_{std}$ over 5 seeds.

	Channel simple, ID	Direct context, ID	Direct instruct, ID	Direct simple, ID	Channel simple, ID mix	Direct context, ID mix	Direct instruct, ID mix	Direct simple, ID mix	Channel simple, OOD mix	Direct context, OOD mix	Direct instruct, OOD mix	Direct simple, OOD mix	0-shot, OOD	1-shot, OOD	4-shot, OOD	8-shot, OOD
SST2	$90.3_{0.2}$	$88.8_{0.3}$	$70.6_{0.2}$	$87.2_{0.3}$	$83.4_{0.3}$	$85.2_{0.2}$	$78.6_{0.2}$	$85.3_{0.2}$	$77.8_{0.4}$	$79.2_{0.3}$	$73.5_{0.4}$	$77.6_{0.5}$	$64.8_{0.0}$	$78.3_{1.6}$	$79.3_{10.8}$	$72.5_{11.1}$
SST2 ADV	$54.7_{0.8}$	$43.1_{1.8}$	$49.1_{1.0}$	$47.2_{1.7}$	$53.6_{1.0}$	$58.5_{0.8}$	$53.1_{0.5}$	$61.1_{1.1}$	$54.3_{1.1}$	$57.4_{0.8}$	$53.0_{0.7}$	$61.2_{0.6}$	$48.7_{0.0}$	$50.4_{2.0}$	$51.8_{6.9}$	$55.8_{7.0}$
MR	$84.2_{0.1}$	$87.3_{0.3}$	$64.2_{0.4}$	$84.6_{0.6}$	$77.5_{0.3}$	$82.8_{0.3}$	$75.4_{0.2}$	$83.1_{0.2}$	$73.0_{0.3}$	$75.6_{0.3}$	$69.7_{0.4}$	$76.6_{0.4}$	$68.2_{0.0}$	$74.9_{2.0}$	$77.4_{8.7}$	$71.1_{9.9}$
CR	$86.2_{0.3}$	$90.5_{0.5}$	$41.2_{0.9}$	$85.4_{0.5}$	$84.6_{0.6}$	$78.0_{0.3}$	$70.6_{0.9}$	$77.8_{0.3}$	$84.7_{0.6}$	$77.2_{0.4}$	$70.3_{0.6}$	$77.6_{0.5}$	$81.7_{0.0}$	$79.1_{76.1}$	$80.1_{5.6}$	$76.2_{6.7}$
Amazon	$90.3_{0.3}$	$90.9_{0.4}$	$61.5_{0.4}$	$79.5_{0.5}$	$87.6_{0.3}$	$85.2_{0.3}$	$86.6_{0.2}$	$85.5_{0.2}$	$87.5_{0.3}$	$85.0_{0.3}$	$86.5_{0.1}$	$85.4_{0.3}$	$83.3_{0.0}$	$86.3_{9.6}$	$83.8_{4.0}$	$80.3_{8.0}$
Yelp	$88.3_{0.1}$	$91.9_{0.2}$	$61.7_{0.3}$	$91.5_{0.2}$	$86.9_{0.1}$	$76.4_{0.3}$	$87.9_{0.1}$	$75.1_{0.3}$	$86.7_{0.2}$	$76.5_{0.3}$	$87.9_{0.2}$	$75.8_{0.3}$	$85.1_{0.0}$	$85.4_{12.6}$	$83.2_{7.1}$	$80.2_{7.7}$
Poem	$82.2_{2.2}$	$78.0_{2.1}$	$48.4_{0.8}$	$76.6_{3.7}$	$80.6_{1.3}$	$85.4_{1.9}$	$70.6_{1.3}$	$79.6_{1.3}$	$82.6_{1.9}$	$84.0_{0.9}$	$70.0_{2.3}$	$79.0_{1.1}$	$57.1_{0.0}$	$70.9_{7.1}$	$68.0_{9.3}$	$52.6_{3.4}$
Finance	$92.2_{1.0}$	$94.7_{1.1}$	$71.8_{1.0}$	$91.9_{0.4}$	$76.0_{0.6}$	$76.6_{0.4}$	$81.9_{0.6}$	$78.0_{0.9}$	$76.1_{0.7}$	$78.3_{0.6}$	$81.4_{0.0}$	$80.6_{0.4}$	$92.3_{0.0}$	$72.9_{8.4}$	$73.6_{18.0}$	$74.3_{32.5}$
SST2 AAVE	$85.3_{0.2}$	$80.1_{0.8}$	$51.9_{0.1}$	$81.1_{0.5}$	$85.3_{0.2}$	$49.2_{0.3}$	$49.1_{0.1}$	$49.4_{0.3}$	$84.3_{0.2}$	$86.2_{0.2}$	$72.7_{0.2}$	$84.9_{0.2}$	$64.2_{0.0}$	$75.3_{3.3}$	$77.6_{9.1}$	$72.3_{11.6}$
SST2 SG	$83.2_{0.3}$	$75.7_{0.6}$	$63.7_{0.5}$	$81.4_{0.6}$	$83.7_{0.1}$	$49.3_{0.3}$	$49.3_{0.2}$	$49.4_{0.3}$	$83.6_{0.1}$	$85.1_{0.3}$	$71.4_{0.2}$	$80.6_{0.4}$	$64.7_{0.0}$	$73.4_{3.0}$	$75.3_{9.7}$	$70.7_{10.8}$

Table 8: Macro-F1 with GPT2-XL. We report $average_{std}$ over 10 runs for zero-shot. For few-shot, we report $average_{std}$ over 5 seeds.

	Channel simple, ID	Direct context, ID	Direct instruct, ID	Direct simple, ID	Channel simple, ID mix	Direct context, ID mix	Direct instruct, ID mix	Direct simple, ID mix	Channel simple, OOD mix	Direct context, OOD mix	Direct instruct, OOD mix	Direct simple, OOD mix	0-shot, OOD	1-shot, OOD	4-shot, OOD	8-shot, OOD
SST2	$88.7_{0.3}$	$89.8_{0.3}$	$87.7_{0.3}$	$86.8_{0.4}$	$75.7_{0.2}$	$87.3_{0.2}$	$82.1_{0.3}$	$84.6_{0.3}$	$69.3_{0.4}$	$88.4_{0.2}$	$80.6_{0.3}$	$85.6_{0.4}$	$67.6_{0.0}$	$83.2_{2.5}$	$88.0_{6.1}$	$87.8_{5.6}$
SST2 ADV	$46.2_{1.2}$	$48.9_{1.1}$	$60.3_{1.1}$	$47.5_{1.2}$	$53.2_{0.8}$	$64.2_{1.1}$	$60.7_{1.1}$	$59.5_{0.9}$	$51.5_{0.9}$	$61.9_{0.9}$	$60.8_{1.2}$	$61.4_{0.7}$	$56.1_{0.0}$	$56.4_{1.7}$	$54.9_{1.9}$	$54.9_{1.5}$
MR	$84.8_{0.2}$	$85.2_{0.3}$	$81.9_{0.5}$	$77.7_{0.4}$	$76.1_{0.3}$	$83.3_{0.3}$	$80.6_{0.3}$	$80.2_{0.3}$	$71.4_{0.2}$	$84.1_{0.3}$	$80.3_{0.2}$	$82.1_{0.3}$	$74.7_{0.0}$	$83.9_{1.4}$	$85.3_{3.9}$	$86.7_{2.5}$
CR	$87.1_{0.3}$	$77.3_{0.6}$	$51.0_{1.6}$	$80.1_{0.8}$	$77.7_{0.6}$	$69.5_{0.5}$	$67.3_{0.6}$	$63.7_{0.5}$	$78.5_{0.8}$	$70.8_{0.7}$	$66.6_{0.6}$	$64.9_{0.6}$	$78.5_{0.0}$	$84.0_{3.1}$	$86.9_{3.0}$	$86.1_{3.7}$
Amazon	$68.6_{0.3}$	$91.2_{0.4}$	$89.3_{0.3}$	$85.1_{0.3}$	$75.1_{0.3}$	$90.2_{0.3}$	$90.1_{0.2}$	$89.2_{0.1}$	$75.1_{0.3}$	$90.5_{0.2}$	$90.3_{0.2}$	$89.3_{0.2}$	$83.9_{0.0}$	$91.6_{0.7}$	$90.9_{1.7}$	$90.6_{2.3}$
Yelp	$84.4_{0.1}$	$87.7_{0.3}$	$73.2_{0.3}$	$74.2_{0.4}$	$84.4_{0.1}$	$87.7_{0.2}$	$86.2_{0.1}$	$84.4_{0.2}$	$84.1_{0.2}$	$88.1_{0.2}$	$85.9_{0.1}$	$84.6_{0.3}$	$93.3_{0.0}$	$94.0_{0.5}$	$92.4_{1.7}$	$92.0_{2.3}$
Poem	$91.0_{1.7}$	$66.0_{4.8}$	$69.2_{1.9}$	$61.4_{2.7}$	$74.6_{1.9}$	$70.4_{1.3}$	$68.4_{0.8}$	$66.2_{1.1}$	$74.6_{1.3}$	$70.6_{1.6}$	$68.4_{0.8}$	$65.8_{1.5}$	$60.0_{0.0}$	$80.6_{2.8}$	$65.1_{11.6}$	$52.6_{3.4}$
Finance	$88.3_{1.0}$	$82.0_{1.4}$	$90.7_{1.3}$	$62.4_{2.6}$	$81.0_{0.6}$	$89.2_{0.7}$	$70.5_{0.5}$	$86.8_{0.6}$	$81.0_{0.7}$	$89.0_{0.4}$	$71.4_{1.0}$	$86.6_{0.3}$	$87.9_{0.0}$	$74.1_{11.7}$	$85.9_{6.9}$	$74.3_{32.5}$
SST2 AAVE	$85.0_{0.3}$	$69.0_{0.7}$	$72.8_{0.6}$	$78.4_{0.5}$	$85.3_{0.2}$	$49.1_{0.4}$	$49.0_{0.4}$	$49.0_{0.3}$	$84.5_{0.2}$	$64.7_{0.5}$	$83.0_{0.3}$	$77.2_{0.4}$	$79.0_{0.0}$	$83.4_{3.0}$	$87.5_{4.6}$	$88.2_{2.6}$
SST2 SG	$82.4_{0.3}$	$62.5_{0.5}$	$57.4_{0.3}$	$72.9_{0.6}$	$83.6_{0.2}$	$49.1_{0.5}$	$49.0_{0.3}$	$49.0_{0.4}$	$83.9_{0.2}$	$70.4_{0.4}$	$83.1_{0.3}$	$78.0_{0.4}$	$76.4_{0.0}$	$81.3_{3.2}$	$86.5_{4.5}$	$87.3_{3.3}$

Table 9: Macro-F1 with OPT 1.3B. We report $average_{std}$ over 10 runs for zero-shot. For few-shot, we report $average_{std}$ over 5 seeds.

	Channel simple, ID	Direct context, ID	Direct instruct, ID	Direct simple, ID	Channel simple, ID mix	Direct context, ID mix	Direct instruct, ID mix	Direct simple, ID mix	Channel simple, OOD mix	Direct context, OOD mix	Direct instruct, OOD mix	Direct simple, OOD mix	0-shot, OOD	1-shot, OOD	4-shot, OOD	8-shot, OOD
SST2	$89.1_{0.2}$	$84.3_{0.6}$	$75.3_{0.7}$	$82.8_{0.5}$	$74.9_{0.2}$	$81.3_{0.3}$	$77.8_{0.1}$	$82.5_{0.3}$	$65.3_{0.4}$	$79.3_{0.3}$	$75.7_{0.2}$	$83.6_{0.3}$	$57.7_{0.0}$	$81.9_{1.5}$	$82.3_{6.0}$	$86.2_{5.3}$
SST2 ADV	$57.0_{1.0}$	$46.1_{1.6}$	$51.1_{2.4}$	$52.4_{1.1}$	$51.4_{0.7}$	$45.9_{0.7}$	$50.5_{0.7}$	$52.2_{0.6}$	$51.7_{0.8}$	$46.1_{0.9}$	$48.9_{0.6}$	$51.2_{0.7}$	$52.0_{0.0}$	$43.2_{3.4}$	$42.0_{4.1}$	$45.5_{2.2}$
MR	$81.3_{0.4}$	$80.7_{0.7}$	$51.9_{0.5}$	$81.2_{0.5}$	$72.5_{0.4}$	$80.3_{0.2}$	$76.9_{0.2}$	$79.4_{0.3}$	$64.6_{0.4}$	$78.5_{0.3}$	$76.7_{0.4}$	$81.4_{0.2}$	$78.4_{0.0}$	$80.2_{3.1}$	$77.9_{6.2}$	$80.7_{9.4}$
CR	$84.5_{0.6}$	$79.2_{0.6}$	$37.2_{0.5}$	$71.3_{1.0}$	$76.2_{0.5}$	$77.3_{0.4}$	$73.6_{0.3}$	$79.6_{0.3}$	$76.9_{0.5}$	$76.6_{0.5}$	$73.5_{0.2}$	$79.1_{0.6}$	$85.4_{0.0}$	$82.6_{6.8}$	$83.6_{3.6}$	$85.7_{2.4}$
Amazon	$88.7_{0.3}$	$90.4_{0.4}$	$80.5_{0.4}$	$89.7_{0.3}$	$81.9_{0.2}$	$88.0_{0.2}$	$84.9_{0.1}$	$89.6_{0.2}$	$82.1_{0.2}$	$87.5_{0.2}$	$84.7_{0.1}$	$89.4_{0.2}$	$86.9_{0.0}$	$87.1_{3.0}$	$85.7_{4.9}$	$88.0_{2.2}$
Yelp	$89.1_{0.2}$	$91.8_{0.2}$	$72.1_{0.4}$	$91.9_{0.2}$	$87.2_{0.3}$	$82.9_{0.2}$	$78.1_{0.1}$	$84.3_{0.1}$	$87.2_{0.2}$	$82.8_{0.2}$	$77.5_{0.1}$	$84.4_{0.1}$	$91.6_{0.0}$	$88.4_{4.5}$	$83.3_{8.5}$	$85.8_{6.1}$
Poem	$87.0_{1.7}$	$79.4_{2.8}$	$62.6_{3.8}$	$81.2_{3.9}$	$79.4_{1.6}$	$75.0_{1.4}$	$54.0_{0.0}$	$75.2_{1.7}$	$76.2_{1.1}$	$73.0_{2.2}$	$54.6_{1.0}$	$73.2_{1.9}$	$54.3_{0.0}$	$81.1_{6.9}$	$69.7_{12.9}$	$52.6_{3.4}$
Finance	$90.4_{0.9}$	$86.7_{1.0}$	$61.7_{1.9}$	$82.9_{1.0}$	$86.2_{0.8}$	$82.6_{0.7}$	$43.4_{0.6}$	$86.6_{0.4}$	$87.7_{1.0}$	$80.9_{0.5}$	$46.1_{0.3}$	$85.4_{0.3}$	$95.3_{0.0}$	$92.8_{1.5}$	$48.3_{25.9}$	$74.3_{32.5}$
SST2 AAVE	$85.1_{0.3}$	$64.7_{0.6}$	$50.0_{0.1}$	$82.4_{0.5}$	$84.9_{0.2}$	$49.1_{0.0}$	$49.1_{0.0}$	$49.1_{0.0}$	$85.0_{0.3}$	$78.5_{0.5}$	$53.2_{0.1}$	$76.7_{0.4}$	$76.3_{0.0}$	$81.7_{1.6}$	$80.0_{6.2}$	$80.5_{10.0}$
SST2 SG	$84.5_{0.4}$	$74.6_{0.7}$	$49.4_{0.1}$	$73.5_{0.6}$	$84.4_{0.2}$	$49.1_{0.0}$	$49.1_{0.0}$	$49.1_{0.0}$	$85.2_{0.2}$	$76.0_{0.3}$	$52.8_{0.2}$	$77.9_{0.3}$	$76.5_{0.0}$	$80.9_{1.1}$	$78.4_{6.3}$	$79.6_{10.0}$

Table 10: Macro-F1 with Pythia 1.4B. We report $average_{std}$ over 10 runs for zero-shot. For few-shot, we report $average_{std}$ over 5 seeds.

	Channel simple, ID	Direct context, ID	Direct instruct, ID	Direct simple, ID	Channel simple, ID mix	Direct context, ID mix	Direct instruct, ID mix	Direct simple, ID mix	Channel simple, OOD mix	Direct context, OOD mix	Direct instruct, OOD mix	Direct simple, OOD mix	0-shot, OOD	1-shot, OOD	4-shot, OOD	8-shot, OOD
SST2	$90.2_{0.2}$	$73.2_{0.5}$	$90.5_{0.3}$	$83.2_{0.6}$	$81.6_{0.4}$	$76.3_{0.4}$	$83.4_{0.2}$	$74.5_{0.4}$	$78.2_{0.3}$	$79.9_{0.5}$	$80.0_{0.4}$	$83.9_{0.4}$	$78.7_{0.0}$	$87.6_{4.5}$	$92.1_{2.7}$	$94.1_{0.9}$
SST2 ADV	$53.7_{1.6}$	$51.3_{2.0}$	$51.1_{1.6}$	$50.5_{1.6}$	$57.5_{0.7}$	$52.4_{1.1}$	$66.8_{0.6}$	$53.3_{0.8}$	$56.4_{1.1}$	$55.2_{1.0}$	$67.4_{0.7}$	$54.9_{1.7}$	$52.0_{0.0}$	$59.9_{4.6}$	$61.9_{7.3}$	$63.2_{4.5}$
MR	$86.4_{0.2}$	$71.4_{0.6}$	$76.3_{0.7}$	$65.3_{0.5}$	$78.2_{0.3}$	$71.2_{0.3}$	$80.0_{0.3}$	$68.8_{0.3}$	$74.9_{0.3}$	$74.9_{0.2}$	$77.2_{0.4}$	$78.5_{0.4}$	$74.7_{0.0}$	$84.5_{3.7}$	$84.9_{6.9}$	$89.6_{1.7}$
CR	$80.3_{0.4}$	$63.6_{1.0}$	$76.1_{1.2}$	$68.6_{1.1}$	$78.0_{0.3}$	$65.6_{0.7}$	$70.5_{0.4}$	$67.3_{0.9}$	$78.3_{0.6}$	$64.0_{1.0}$	$66.5_{0.5}$	$65.5_{0.8}$	$80.6_{0.0}$	$85.3_{2.7}$	$87.9_{2.9}$	$88.1_{1.3}$
Amazon	$86.0_{0.3}$	$60.7_{0.4}$	$77.2_{0.6}$	$62.4_{0.4}$	$87.4_{0.2}$	$82.3_{0.3}$	$89.8_{0.2}$	$83.5_{0.2}$	$87.4_{0.2}$	$83.1_{0.3}$	$90.8_{0.2}$	$84.0_{0.3}$	$83.8_{0.0}$	$90.8_{1.2}$	$88.1_{4.3}$	$87.8_{4.3}$
Yelp	$89.0_{0.2}$	$74.9_{0.5}$	$74.6_{0.4}$	$65.9_{0.5}$	$85.9_{0.1}$	$66.6_{0.3}$	$86.3_{0.1}$	$67.8_{0.2}$	$86.1_{0.2}$	$68.7_{0.1}$	$87.3_{0.1}$	$70.0_{0.2}$	$90.3_{0.0}$	$90.8_{1.9}$	$86.8_{4.5}$	$86.1_{5.1}$
Poem	$88.4_{2.5}$	$63.8_{2.9}$	$71.4_{2.3}$	$71.8_{2.9}$	$70.2_{1.1}$	$71.4_{1.0}$	$91.0_{1.1}$	$70.0_{1.6}$	$73.0_{1.1}$	$72.6_{1.3}$	$88.6_{1.3}$	$69.6_{2.1}$	$82.9_{0.0}$	$77.7_{11.2}$	$64.6_{3.9}$	$52.6_{3.4}$
Finance	$91.6_{0.4}$	$77.5_{2.3}$	$49.3_{1.2}$	$64.3_{2.3}$	$81.9_{0.6}$	$88.4_{0.7}$	$75.5_{1.5}$	$87.4_{0.8}$	$83.2_{0.7}$	$88.6_{0.3}$	$71.7_{1.3}$	$87.3_{0.3}$	$90.2_{0.0}$	$82.0_{13.9}$	$78.7_{21.0}$	$74.3_{32.5}$
SST2 AAVE	$86.9_{0.3}$	$77.4_{0.6}$	$60.6_{0.2}$	$79.4_{0.7}$	$87.0_{0.3}$	$49.1_{0.0}$	$49.1_{0.0}$	$49.1_{0.0}$	$87.3_{0.4}$	$71.5_{0.5}$	$69.9_{0.4}$	$77.5_{0.4}$	$79.8_{0.0}$	$85.6_{4.7}$	$86.9_{7.2}$	$92.0_{1.2}$
SST2 SG	$85.4_{0.3}$	$69.4_{0.4}$	$52.4_{0.3}$	$74.4_{0.6}$	$83.5_{0.2}$	$49.1_{0.0}$	$49.1_{0.0}$	$49.1_{0.0}$	$83.0_{0.2}$	$74.6_{0.3}$	$73.7_{0.3}$	$76.7_{0.5}$	$79.7_{0.0}$	$84.5_{5.2}$	$85.3_{7.0}$	$91.4_{1.6}$

Table 11: Macro-F1 with OPT 2.7B. We report $average_{std}$ over 10 runs for zero-shot. For few-shot, we report $average_{std}$ over 5 seeds. Further results are provided in the appendix

	Channel simple, ID	Direct context, ID	Direct instruct, ID	Direct simple, ID	Channel simple, ID mix	Direct context, ID mix	Direct instruct, ID mix	Direct simple, ID mix	Channel simple, OOD mix	Direct context, OOD mix	Direct instruct, OOD mix	Direct simple, OOD mix	0-shot, OOD	1-shot, OOD	4-shot, OOD	8-shot, OOD
SST2	$91.4_{0.2}$	$79.9_{0.6}$	$88.5_{0.3}$	$85.4_{0.5}$	$80.1_{0.4}$	$81.9_{0.3}$	$77.6_{0.4}$	$82.8_{0.5}$	$71.4_{0.3}$	$70.4_{0.3}$	$78.1_{0.5}$	$69.3_{0.4}$	$60.2_{0.0}$	$87.3_{1.4}$	$90.3_{1.5}$	$91.0_{1.5}$
SST2 ADV	$47.8_{0.9}$	$53.7_{1.8}$	$58.9_{2.0}$	$51.3_{1.8}$	$47.7_{0.7}$	$55.7_{1.0}$	$55.9_{1.2}$	$54.3_{0.8}$	$47.7_{0.8}$	$57.1_{1.0}$	$57.3_{0.9}$	$53.4_{0.9}$	$45.3_{0.0}$	$49.2_{2.5}$	$52.7_{2.1}$	$52.7_{1.7}$
MR	$86.0_{0.3}$	$80.3_{0.5}$	$72.8_{0.5}$	$79.1_{0.7}$	$74.9_{0.3}$	$76.5_{0.3}$	$74.6_{0.3}$	$77.6_{0.3}$	$67.6_{0.4}$	$66.3_{0.3}$	$76.5_{0.3}$	$64.2_{0.4}$	$73.5_{0.0}$	$84.6_{2.1}$	$87.0_{0.9}$	$87.0_{2.3}$
CR	$86.1_{0.4}$	$78.7_{0.9}$	$52.3_{0.8}$	$79.0_{1.1}$	$85.3_{0.2}$	$71.9_{0.4}$	$70.7_{0.4}$	$60.9_{0.9}$	$84.0_{0.3}$	$71.8_{0.4}$	$71.0_{0.4}$	$62.9_{0.9}$	$76.9_{0.0}$	$85.9_{3.1}$	$86.2_{3.5}$	$83.3_{4.2}$
Amazon	$89.2_{0.4}$	$72.7_{0.6}$	$80.1_{0.7}$	$71.2_{0.6}$	$87.4_{0.3}$	$86.1_{0.3}$	$81.0_{0.2}$	$84.0_{0.2}$	$87.6_{0.2}$	$82.6_{0.4}$	$84.1_{0.2}$	$81.7_{0.3}$	$85.6_{0.0}$	$88.3_{3.0}$	$91.8_{2.1}$	$90.2_{4.3}$
Yelp	$87.0_{0.3}$	$90.0_{0.2}$	$81.2_{0.2}$	$85.6_{0.4}$	$86.9_{0.2}$	$68.1_{0.2}$	$66.5_{0.3}$	$64.9_{0.2}$	$86.8_{0.2}$	$67.8_{0.2}$	$66.4_{0.1}$	$64.4_{0.2}$	$87.8_{0.0}$	$87.5_{4.7}$	$93.4_{2.3}$	$85.7_{9.1}$
Poem	$84.6_{1.6}$	$80.2_{2.6}$	$53.2_{3.0}$	$73.6_{2.5}$	$95.4_{1.9}$	$70.0_{1.6}$	$71.0_{1.9}$	$58.4_{0.8}$	$94.4_{1.8}$	$69.4_{1.9}$	$66.0_{2.5}$	$55.0_{1.4}$	$60.0_{0.0}$	$76.6_{14.2}$	$58.3_{5.0}$	$52.6_{3.4}$
Finance	$93.8_{0.6}$	$79.9_{1.9}$	$37.8_{1.3}$	$73.8_{1.3}$	$86.3_{0.6}$	$81.2_{0.6}$	$68.4_{0.3}$	$78.9_{0.5}$	$85.5_{0.6}$	$78.9_{0.7}$	$67.9_{0.8}$	$76.9_{1.0}$	$58.3_{0.0}$	$94.3_{2.3}$	$88.2_{4.4}$	$74.3_{32.5}$
SST2 AAVE	$88.0_{0.3}$	$71.9_{0.9}$	$59.8_{0.5}$	$70.2_{1.2}$	$87.6_{0.1}$	$49.1_{0.0}$	$49.1_{0.0}$	$49.1_{0.0}$	$89.1_{0.3}$	$71.0_{0.5}$	$79.3_{0.3}$	$69.8_{0.4}$	$73.1_{0.0}$	$86.1_{1.9}$	$87.3_{1.5}$	$87.9_{2.8}$
SST2 SG	$87.1_{0.4}$	$66.6_{0.7}$	$54.5_{0.2}$	$61.5_{0.9}$	$86.1_{0.2}$	$49.1_{0.0}$	$49.1_{0.0}$	$49.1_{0.0}$	$85.1_{0.2}$	$73.0_{0.5}$	$73.1_{0.3}$	$68.1_{0.5}$	$72.3_{0.0}$	$84.5_{2.6}$	$85.0_{2.3}$	$86.3_{3.6}$

Table 12: Macro-F1 with Pythia 2.8B. We report $average_{std}$ over 10 runs for zero-shot. For few-shot, we report $average_{std}$ over 5 seeds.

	Channel simple, ID	Direct context, ID	Direct instruct, ID	Direct simple, ID	Channel simple, ID mix	Direct context, ID mix	Direct instruct, ID mix	Direct simple, ID mix	Channel simple, OOD mix	Direct context, OOD mix	Direct instruct, OOD mix	Direct simple, OOD mix	0-shot, OOD	1-shot, OOD	4-shot, OOD	8-shot, OOD
SST2	$90.7_{0.2}$	$86.7_{0.3}$	$49.9_{0.2}$	$82.0_{0.5}$	$80.6_{0.3}$	$84.4_{0.2}$	$73.2_{0.2}$	$83.3_{0.3}$	$72.2_{0.4}$	$82.8_{0.5}$	$68.8_{0.3}$	$81.9_{0.4}$	$50.7_{0.0}$	$88.9_{3.2}$	$93.6_{0.8}$	$93.6_{1.3}$
SST2 ADV	$52.5_{1.2}$	$57.1_{1.7}$	$55.3_{1.2}$	$55.4_{1.6}$	$53.4_{0.6}$	$53.5_{1.0}$	$59.2_{0.7}$	$60.3_{1.2}$	$53.0_{0.9}$	$60.2_{0.8}$	$58.3_{0.5}$	$56.2_{1.6}$	$47.3_{0.0}$	$52.4_{5.6}$	$58.2_{3.7}$	$56.4_{6.1}$
MR	$84.7_{0.3}$	$75.9_{0.6}$	$53.2_{0.3}$	$65.9_{1.0}$	$77.3_{0.4}$	$80.4_{0.1}$	$71.8_{0.3}$	$80.9_{0.2}$	$69.7_{0.4}$	$80.5_{0.2}$	$67.4_{0.2}$	$80.4_{0.3}$	$55.3_{0.0}$	$85.7_{2.2}$	$90.2_{0.7}$	$89.5_{1.4}$
CR	$86.4_{0.6}$	$68.6_{0.9}$	$36.9_{0.2}$	$60.4_{1.2}$	$86.5_{0.4}$	$70.6_{0.5}$	$68.1_{0.6}$	$66.1_{0.8}$	$86.0_{0.2}$	$67.7_{0.8}$	$56.9_{0.4}$	$65.9_{0.5}$	$68.4_{0.0}$	$87.7_{0.8}$	$87.3_{1.8}$	$86.8_{3.1}$
Amazon	$89.3_{0.3}$	$82.2_{0.5}$	$65.5_{0.5}$	$74.2_{0.5}$	$87.6_{0.2}$	$50.1_{0.3}$	$50.1_{0.3}$	$50.1_{0.3}$	$86.9_{0.2}$	$85.5_{0.2}$	$89.7_{0.1}$	$84.6_{0.2}$	$83.8_{0.0}$	$93.6_{0.1}$	$93.7_{0.3}$	$93.5_{0.2}$
Yelp	$89.8_{0.2}$	$75.3_{0.3}$	$66.0_{0.4}$	$76.8_{0.2}$	$85.9_{0.1}$	$50.0_{0.2}$	$50.0_{0.2}$	$50.0_{0.2}$	$85.4_{0.2}$	$84.1_{0.3}$	$50.2_{0.2}$	$82.6_{0.3}$	$95.1_{0.0}$	$96.2_{0.5}$	$95.5_{0.4}$	$94.2_{0.6}$
Poem	$96.0_{1.3}$	$72.0_{2.5}$	$62.0_{3.1}$	$68.8_{3.6}$	$88.8_{1.0}$	$78.8_{1.4}$	$63.8_{1.1}$	$71.0_{1.1}$	$88.0_{1.3}$	$76.8_{1.4}$	$51.4_{1.6}$	$75.2_{1.9}$	$54.3_{0.0}$	$85.7_{13.3}$	$80.0_{9.9}$	$52.6_{3.4}$
Finance	$92.8_{0.8}$	$79.3_{0.7}$	$71.8_{1.4}$	$85.0_{0.7}$	$87.7_{1.0}$	$83.4_{1.0}$	$84.6_{0.9}$	$73.3_{0.5}$	$87.1_{0.6}$	$89.5_{0.8}$	$92.9_{1.2}$	$78.7_{0.5}$	$92.3_{0.0}$	$95.1_{0.6}$	$90.4_{5.0}$	$74.3_{32.5}$
SST2 AAVE	$87.0_{0.4}$	$70.3_{0.6}$	$49.1_{0.1}$	$81.2_{0.2}$	$87.6_{0.2}$	$51.2_{0.5}$	$49.5_{0.2}$	$55.2_{0.5}$	$87.5_{0.2}$	$80.9_{0.4}$	$49.2_{0.1}$	$76.4_{0.2}$	$52.9_{0.0}$	$88.0_{2.4}$	$90.8_{2.1}$	$90.4_{2.8}$
SST2 SG	$86.5_{0.2}$	$90.1_{0.3}$	$52.7_{0.7}$	$73.2_{0.5}$	$86.7_{0.2}$	$50.6_{0.4}$	$50.1_{0.2}$	$49.1_{0.0}$	$86.9_{0.2}$	$69.2_{0.3}$	$53.2_{0.3}$	$73.4_{0.4}$	$52.5_{0.0}$	$86.4_{2.6}$	$89.3_{1.8}$	$88.9_{2.9}$

Table 13: Macro-F1 with Pythia 6.9B. We report $average_{std}$ over 10 runs for zero-shot. For few-shot, we report $average_{std}$ over 5 seeds.

	Channel simple, ID	Direct context, ID	Direct instruct, ID	Direct simple, ID	Channel simple, ID mix	Direct context, ID mix	Direct instruct, ID mix	Direct simple, ID mix	Channel simple, OOD mix	Direct context, OOD mix	Direct instruct, OOD mix	Direct simple, OOD mix	0-shot, OOD	1-shot, OOD	4-shot, OOD	8-shot, OOD
SST2	$92.1_{0.3}$	$82.1_{0.5}$	$68.9_{0.4}$	$80.3_{1.1}$	$90.8_{0.2}$	$85.0_{0.4}$	$76.2_{0.4}$	$85.2_{0.3}$	$85.6_{0.3}$	$70.2_{0.4}$	$57.0_{0.3}$	$64.6_{0.6}$	$89.3_{0.0}$	$61.1_{10.0}$	$71.8_{11.3}$	$70.6_{10.5}$
SST2 ADV	$59.2_{0.8}$	$61.6_{1.3}$	$56.1_{1.7}$	$62.2_{2.2}$	$73.9_{0.6}$	$58.0_{0.7}$	$66.5_{0.6}$	$64.4_{0.8}$	$74.1_{0.8}$	$60.0_{0.8}$	$64.2_{1.0}$	$65.7_{0.9}$	$73.0_{0.0}$	$61.0_{5.2}$	$62.7_{5.8}$	$61.0_{5.6}$
MR	$90.2_{0.2}$	$74.8_{0.7}$	$63.6_{0.5}$	$56.1_{0.7}$	$86.4_{0.2}$	$78.0_{0.3}$	$74.9_{0.3}$	$80.8_{0.2}$	$81.3_{0.2}$	$65.1_{0.3}$	$57.8_{0.4}$	$60.7_{0.4}$	$86.9_{0.0}$	$56.7_{6.9}$	$70.2_{10.7}$	$61.6_{6.9}$
CR	$86.1_{0.3}$	$72.0_{1.2}$	$50.8_{1.5}$	$73.4_{1.6}$	$87.8_{0.4}$	$73.9_{0.7}$	$76.8_{1.0}$	$69.2_{0.6}$	$87.5_{0.4}$	$76.4_{0.6}$	$76.2_{0.6}$	$70.7_{0.8}$	$51.1_{0.0}$	$48.0_{9.8}$	$67.8_{8.1}$	$71.5_{8.1}$
Amazon	$88.4_{0.4}$	$92.3_{0.3}$	$87.5_{0.2}$	$86.2_{0.3}$	$91.2_{0.2}$	$93.0_{0.2}$	$80.1_{0.3}$	$89.6_{0.2}$	$91.4_{0.3}$	$92.3_{0.1}$	$79.7_{0.2}$	$89.2_{0.2}$	$89.7_{0.0}$	$59.6_{10.4}$	$74.0_{7.9}$	$82.9_{5.1}$
Yelp	$93.7_{0.1}$	$87.2_{0.2}$	$90.4_{0.3}$	$77.1_{0.4}$	$92.4_{0.1}$	$92.7_{0.1}$	$91.0_{0.2}$	$90.0_{0.2}$	$92.3_{0.1}$	$92.8_{0.1}$	$92.2_{0.1}$	$89.6_{0.2}$	$84.7_{0.0}$	$66.1_{8.6}$	$72.6_{9.2}$	$82.5_{6.8}$
Poem	$91.6_{0.8}$	$89.2_{1.0}$	$84.4_{1.8}$	$82.6_{1.6}$	$89.6_{0.8}$	$82.0_{0.9}$	$67.2_{2.7}$	$74.0_{1.6}$	$90.0_{0.0}$	$82.0_{0.0}$	$66.0_{2.7}$	$73.0_{1.4}$	$68.6_{0.0}$	$58.3_{6.4}$	$59.4_{6.9}$	$54.3_{0.0}$
Finance	$96.4_{0.0}$	$97.1_{0.6}$	$46.1_{2.6}$	$81.3_{0.6}$	$91.3_{0.9}$	$81.9_{0.5}$	$60.2_{0.9}$	$69.8_{0.5}$	$90.7_{0.6}$	$82.2_{1.2}$	$60.7_{0.6}$	$69.4_{1.2}$	$58.6_{0.0}$	$86.5_{4.4}$	$73.7_{32.1}$	$90.6_{0.0}$
SST2 AAVE	$88.8_{0.3}$	$65.3_{0.6}$	$63.8_{0.5}$	$60.2_{0.4}$	$89.1_{0.1}$	$49.1_{0.0}$	$49.1_{0.0}$	$49.1_{0.0}$	$89.3_{0.2}$	$58.2_{0.4}$	$53.1_{0.3}$	$57.4_{0.4}$	$86.4_{0.0}$	$59.7_{7.8}$	$71.4_{10.1}$	$63.6_{10.1}$
SST2 SG	$87.3_{0.2}$	$56.4_{0.4}$	$50.3_{0.3}$	$55.2_{0.3}$	$88.6_{0.2}$	$49.1_{0.0}$	$49.1_{0.0}$	$49.1_{0.0}$	$88.6_{0.3}$	$62.2_{0.5}$	$63.0_{0.4}$	$62.2_{0.4}$	$84.5_{0.0}$	$61.1_{8.1}$	$69.3_{9.1}$	$65.2_{11.5}$

Table 14: Macro-F1 with Mistral 7B. We report $average_{std}$ over 10 runs for zero-shot. For few-shot, we report $average_{std}$ over 5 seeds. Further results are provided in the appendix

	Channel simple, ID	Direct context, ID	Direct instruct, ID	Direct simple, ID	Channel simple, ID mix	Direct context, ID mix	Direct instruct, ID mix	Direct simple, ID mix	Channel simple, OOD mix	Direct context, OOD mix	Direct instruct, OOD mix	Direct simple, OOD mix	0-shot, OOD	1-shot, OOD	4-shot, OOD	8-shot, OOD
SST2	$90.5_{0.3}$	$56.5_{1.0}$	$49.7_{0.8}$	$53.9_{0.6}$	$81.2_{0.3}$	$79.8_{0.4}$	$52.1_{0.2}$	$75.3_{0.5}$	$73.1_{0.3}$	$73.1_{0.3}$	$49.7_{0.1}$	$73.1_{0.4}$	$68.9_{0.0}$	$49.8_{1.0}$	$56.4_{13.4}$	$60.8_{13.6}$
SST2 ADV	$54.7_{1.7}$	$57.9_{1.0}$	$59.3_{2.0}$	$60.4_{0.9}$	$52.8_{0.7}$	$62.2_{1.4}$	$50.1_{0.3}$	$64.5_{1.3}$	$53.0_{0.8}$	$58.8_{1.4}$	$49.3_{0.5}$	$61.1_{1.2}$	$52.7_{0.0}$	$54.2_{3.5}$	$52.7_{3.5}$	$54.3_{3.3}$
MR	$88.6_{0.2}$	$58.5_{0.9}$	$51.6_{0.3}$	$60.6_{1.1}$	$76.3_{0.4}$	$74.2_{0.4}$	$52.9_{0.2}$	$73.0_{0.3}$	$69.5_{0.3}$	$68.9_{0.4}$	$51.4_{0.4}$	$71.5_{0.2}$	$67.6_{0.0}$	$50.6_{0.4}$	$55.5_{9.2}$	$56.2_{8.9}$
CR	$82.7_{0.6}$	$56.9_{1.6}$	$60.7_{1.3}$	$53.9_{1.3}$	$82.6_{0.3}$	$63.5_{0.7}$	$37.1_{0.2}$	$62.4_{0.5}$	$82.7_{0.4}$	$61.5_{0.8}$	$37.2_{0.2}$	$58.8_{0.9}$	$49.7_{0.0}$	$48.5_{9.3}$	$51.9_{8.7}$	$53.5_{13.4}$
Amazon	$79.5_{0.2}$	$53.5_{0.6}$	$62.5_{0.6}$	$55.3_{0.9}$	$87.5_{0.2}$	$82.0_{0.4}$	$57.5_{0.4}$	$80.5_{0.4}$	$87.6_{0.2}$	$82.2_{0.4}$	$57.6_{0.3}$	$80.1_{0.3}$	$69.1_{0.0}$	$51.2_{0.7}$	$66.6_{3.9}$	$62.1_{7.2}$
Yelp	$90.4_{0.3}$	$59.9_{0.4}$	$67.5_{0.5}$	$60.9_{0.3}$	$90.6_{0.2}$	$74.0_{0.2}$	$52.9_{0.2}$	$69.1_{0.3}$	$90.5_{0.1}$	$74.0_{0.3}$	$52.9_{0.1}$	$68.0_{0.2}$	$60.1_{0.0}$	$50.4_{0.5}$	$63.5_{6.7}$	$62.2_{6.0}$
Poem	$76.4_{1.6}$	$53.4_{6.3}$	$51.8_{5.5}$	$50.8_{3.4}$	$85.6_{1.6}$	$73.6_{2.3}$	$48.0_{0.0}$	$64.6_{1.0}$	$86.0_{2.3}$	$71.0_{2.4}$	$48.0_{0.0}$	$64.6_{1.6}$	$80.0_{0.0}$	$51.4_{3.6}$	$58.9_{6.9}$	$54.3_{0.0}$
Finance	$94.1_{0.4}$	$88.3_{3.3}$	$78.6_{1.7}$	$87.7_{2.2}$	$93.4_{0.4}$	$68.7_{0.5}$	$43.8_{0.3}$	$69.5_{0.4}$	$93.7_{0.4}$	$68.0_{0.7}$	$44.2_{0.4}$	$69.0_{0.5}$	$90.9_{0.0}$	$73.7_{32.2}$	$72.0_{18.6}$	$90.6_{0.0}$
SST2 AAVE	$82.7_{0.4}$	$72.9_{0.9}$	$54.9_{1.1}$	$69.8_{0.8}$	$82.9_{0.2}$	$49.1_{0.0}$	$49.1_{0.0}$	$49.1_{0.0}$	$85.0_{0.2}$	$58.3_{0.4}$	$49.1_{0.0}$	$51.2_{0.4}$	$66.6_{0.0}$	$50.9_{1.4}$	$54.4_{10.0}$	$56.3_{11.7}$
SST2 SG	$84.1_{0.3}$	$56.2_{1.0}$	$50.1_{0.5}$	$51.2_{1.2}$	$85.8_{0.1}$	$49.1_{0.0}$	$49.1_{0.0}$	$49.1_{0.0}$	$86.1_{0.2}$	$65.3_{0.6}$	$53.7_{0.3}$	$70.1_{0.5}$	$60.1_{0.0}$	$50.9_{1.4}$	$54.4_{9.6}$	$56.2_{12.0}$

Table 15: Macro-F1 with Llama 7B. We report $average_{std}$ over 10 runs for zero-shot. For few-shot, we report $average_{std}$ over 5 seeds. Further results are provided in the appendix

	Channel simple, ID	Direct context, ID	Direct instruct, ID	Direct simple, ID	Channel simple, ID mix	Direct context, ID mix	Direct instruct, ID mix	Direct simple, ID mix	Channel simple, OOD mix	Direct context, OOD mix	Direct instruct, OOD mix	Direct simple, OOD mix	0-shot, OOD	1-shot, OOD	4-shot, OOD	8-shot, OOD
SST2	$92.5_{0.2}$	$86.4_{0.4}$	$55.8_{0.4}$	$85.3_{0.4}$	$81.8_{0.4}$	$82.6_{0.6}$	$59.8_{0.1}$	$79.0_{0.2}$	$73.5_{0.3}$	$81.9_{0.2}$	55.60.1	$82.7_{0.3}$	$64.7_{0.0}$	$56.8_{15.1}$	$65.6_{11.8}$	$69.0_{8.9}$
SST2 ADV	$70.9_{1.1}$	$51.2_{0.8}$	$50.1_{1.1}$	$49.3_{0.7}$	$60.8_{0.5}$	$58.1_{0.8}$	$51.4_{0.0}$	$59.7_{1.3}$	$60.1_{0.8}$	$56.1_{1.5}$	$52.0_{0.0}$	$59.3_{1.2}$	$58.1_{0.0}$	$55.5_{4.8}$	$56.8_{4.1}$	$54.7_{3.4}$
MR	$88.8_{0.1}$	$80.3_{0.9}$	$64.3_{0.6}$	$63.9_{0.6}$	$77.3_{0.3}$	$79.5_{0.3}$	$60.6_{0.2}$	$75.1_{0.5}$	$71.4_{0.3}$	$80.3_{0.2}$	$55.5_{0.2}$	$80.1_{0.2}$	$56.9_{0.0}$	$56.6_{12.8}$	$58.7_{7.8}$	$58.7_{9.3}$
CR	$90.3_{0.3}$	$73.6_{1.3}$	$37.0_{0.4}$	$69.1_{0.8}$	$86.9_{0.5}$	$62.7_{0.8}$	$39.1_{0.1}$	$60.2_{0.6}$	$86.6_{0.4}$	$59.7_{0.6}$	$39.4_{0.0}$	$60.1_{0.5}$	$61.7_{0.0}$	$42.8_{12.3}$	$52.8_{7.5}$	$62.0_{10.0}$
Amazon	$90.7_{0.1}$	$89.7_{0.3}$	$60.2_{1.0}$	$84.5_{0.4}$	$88.5_{0.3}$	$88.3_{0.2}$	$65.8_{0.2}$	$83.7_{0.2}$	$88.2_{0.2}$	$88.5_{0.2}$	$65.8_{0.2}$	$83.8_{0.3}$	$83.0_{0.0}$	$56.6_{1.7}$	$70.8_{7.7}$	$73.6_{4.7}$
Yelp	$93.0_{0.1}$	$93.5_{0.3}$	$72.1_{0.5}$	$69.4_{0.3}$	$89.3_{0.1}$	$64.7_{0.2}$	$70.3_{0.1}$	$61.0_{0.2}$	$89.2_{0.1}$	$64.8_{0.1}$	$70.1_{0.1}$	$60.1_{0.2}$	$54.8_{0.0}$	$53.0_{2.9}$	$64.7_{15.7}$	$67.9_{11.8}$
Poem	$81.2_{1.0}$	$89.0_{1.7}$	$79.8_{2.0}$	$87.0_{3.2}$	$81.2_{1.0}$	$77.8_{0.6}$	$50.0_{0.0}$	$62.2_{1.1}$	$81.4_{1.9}$	$79.2_{1.9}$	$48.4_{0.8}$	$63.4_{1.9}$	$60.0_{0.0}$	$54.9_{1.1}$	$53.7_{6.9}$	$54.3_{0.0}$
Finance	$96.3_{0.3}$	$97.2_{0.6}$	$80.8_{0.8}$	$92.8_{1.1}$	$91.6_{0.6}$	$73.3_{0.3}$	$47.7_{0.6}$	$69.7_{0.6}$	$91.6_{0.7}$	$73.5_{0.6}$	$50.0_{0.6}$	$69.8_{0.6}$	$88.9_{0.0}$	$74.8_{11.5}$	$63.0_{30.8}$	$90.6_{0.0}$
SST2 AAVE	$88.5_{0.4}$	$79.8_{0.5}$	$49.3_{0.1}$	$76.3_{0.6}$	$89.2_{0.2}$	$49.1_{0.0}$	$49.1_{0.0}$	$49.1_{0.0}$	$89.1_{0.1}$	$74.9_{0.1}$	$71.0_{0.7}$	$75.6_{0.4}$	$57.2_{0.0}$	$56.2_{13.2}$	$58.5_{8.6}$	$59.4_{9.7}$
SST2 SG	$86.5_{0.2}$	$90.1_{0.3}$	$52.7_{0.7}$	$73.2_{0.5}$	$86.7_{0.2}$	$50.6_{0.4}$	$50.1_{0.2}$	$49.1_{0.0}$	$86.9_{0.2}$	$69.2_{0.3}$	$53.2_{0.3}$	$73.4_{0.4}$	$54.8_{0.0}$	$50.1_{3.1}$	$49.0_{11.6}$	$56.1_{8.5}$

Table 16: Macro-F1 with Llama-2-7B. We report $average_{std}$ over 10 runs for zero-shot. For few-shot, we report $average_{std}$ over 5 seeds.

Task (Distribution Shift)	Label Descriptions
Sentiment (Domain)	"This [DOMAIN] leans [POLARITY]:"; DOMAIN $\in$ {text, movie review, RottenTomatoes review, customer review, Amazon review, Yelp review, poem, financial news excerpt}, POLARITY $\in$ {positive, negative}
Sentiment (Adversarial)	"This misleading text exhibits a [POLARITY] bias:"; POLARITY $\in$ {positive, negative}
Sentiment (Dialect)	"This movie review written in [DIALECT] leans [POLARITY]:"; DIALECT $\in$ {African American Vernacular English, Colloquial Singapore English (Singlish), Indian English, Irish English, Nigerian English}, POLARITY $\in$ {positive, negative}
Emotion (Dialect)	"This tweet written in [DIALECT] emotes [EMOTION]:"; DIALECT $\in$ {African American Vernacular English, Colloquial Singapore English (Singlish), Indian English, Irish English, Nigerian English}, EMOTION $\in$ {joy, sadness, anger, fear, love, surprise}
News (Dialect)	"This news written in [DIALECT] is about [TOPIC]:"; DIALECT $\in$ {African American Vernacular English, Colloquial Singapore English (Singlish), Indian English, Irish English, Nigerian English}, TOPIC $\in \{world, sports, business, technology\}$

Table 17: Handwritten label description starter templates. DOMAIN="text" represents missing domain information. We generate label description variations by asking ChatGPT "Write 9 paraphrases of this sentence as a Python list"

Name	Format
Direct context	"This is a [DISTRIBUTIONAL INFO]. [INPUT] [LABEL DESCRIPTION]"
Direct instruct	"Is this [DISTRIBUTIONAL INFO] [LIST OF LABEL NAMES]? [INPUT] [LABEL DESCRIPTION]"
Direct simple	"[INPUT] [LABEL DESCRIPTION]"
Attribute	**Example**
DISTRIBUTIONAL INFO	poem
INPUT	with pale blue berries. in these peaceful shades-
LABEL DESCRIPTION	positive
LIST OF LABEL NAMES	negative or positive

Table 18: Different variations of discriminative (direct) models. In all three settings the label descriptions contain contextual information.