Effective In-Context Example Selection through Data Compression

Zhongxiang Sun⋆ and **Kepu Zhang**⋆ and **Haoyu Wang** and **Xiao Zhang** and **Jun Xu**
Gaoling School of Artificial Intelligence,
Renmin University of China
{sunzhongxiang, kepuzhang, wanghaoyu0924, zhangx89, junxu}@ruc.edu.cn

Abstract

In-context learning has been extensively validated in large language models. However, the mechanism and selection strategy for in-context example selection, which is a crucial ingredient in this approach, lacks systematic and in-depth research. In this paper, we propose a data compression approach to the selection of in-context examples. We introduce a two-stage method that can effectively choose relevant examples and retain sufficient information about the training dataset within the in-context examples. Our method shows a significant improvement of an average of 5.90% across five different real-world datasets using four language models.

1 Introduction

Drawing inspiration from recent research that regards Large Language Models (LLMs) as an efficient means of compressing pre-training datasets, and the notion that In-Context Learning (ICL) can be seen as fine-tuning on example datasets (Dai et al., 2022), we assume that LLMs can achieve data re-compression through ICL. In other words, an effective training dataset compression method can aid in the selection of in-context examples. Looking at the matter from another perspective, it is evident that fine-tuning the entire dataset would yield the best results. However, in the case of ICL, we typically choose only a few examples as LLM prompts due to the limitations of input window length. By employing data compression techniques, we can ensure that the majority of data information is preserved in the in-context examples, which is also the aim of dataset pruning.

Based on the aforementioned analysis, we propose to utilize the influence function (Koh and Liang, 2017), which has exhibited efficacy in dataset pruning, to select examples for ICL. However, recent studies on ICL have revealed that the relevance between the query source and in-context examples is critical for ICL. Furthermore, the influence function requires the gradient of parameters, which is computationally expensive and inefficient. To tackle the aforementioned issues, we suggest a two-stage method. Firstly, relevant examples for the query input are recalled, which ensures the correlation between the examples and the query source. Secondly, our meta-gradient-based influence function is utilized to calculate the influence score for each recalled example. Finally, based on the influence score, in-context examples are selected from the recalled examples. Notably, our framework compresses important information from the training set into the in-context examples, thereby enhancing the performance of ICL. Additionally, our framework is data-independent, relies solely on a small number of model parameters, and does not require the training of any additional models. Numerous experiments indicate that our method shows a significant improvement of an average of 5.90% on five different real-world datasets using multiple language models.

⋆ Equal contribution

2 Background

2.1 In-Context Learning

The ICL scenario of LLMs can be regarded as a conditional text generation problem. Concretely, the probability of generating a target y is conditioned on the context C, which includes k examples and the source x. Therefore, the probability can be expressed as:

$$p_{\text{LLM}}(y \mid C, x) = \prod_{t=1}^{T} p\left(y_t \mid C, x, y_{<t}\right)$$

where LLM denotes the parameters of the large language model, and $C = \{x_1, y_1, x_2, y_2, \ldots, x_k, y_k\}$ is a context string concatenating k training instances. For example,

(x_k, y_k) is concatenated with the special character, e.g., "\n" or "Sentence: x; Sentiment y." which is denoted as p_k. In this paper, we have different example sets C at different stages, C_1 in the first stage, and C_2 in the second stage, where C_2 is a subset of C_1.

Dai et al. (2022) explains language models as meta-optimizers and understands ICL as a kind of implicit finetuning:

$$\begin{aligned}\widetilde{\mathcal{F}}_{\mathrm{ICL}}(\mathbf{q}) &= W_{\mathrm{ZSL}}\mathbf{q} + \sum_i \left(W_V\mathbf{x}'_i \otimes (W_K\mathbf{x}'_i)^T\right)\mathbf{q} \\ &= W_{\mathrm{ZSL}}\mathbf{q} + \Delta W_{\mathrm{ICL}}\mathbf{q} \\ &= (W_{\mathrm{ZSL}} + \Delta W_{\mathrm{ICL}})\,\mathbf{q},\end{aligned} \tag{1}$$

where ZSL denotes the zero-shot learning, which only contains the source x; $\mathbf{x} \in \mathbb{R}^d$ is the input representation of a query token t, and $\mathbf{q} = W_Q\mathbf{x} \in \mathbb{R}^{d'}$ is the attention query vector; $\mathbf{x}' \in \mathbb{R}^d$ denotes the input representations of the example's token; $W_Q, W_K, W_V \in \mathbb{R}^{d' \times d}$ are the projection matrices for computing the attention queries, keys, and values, respectively. Dai et al. (2022) regards $W_V X'$ as some meta-gradients, which are used to compute the updated matrix ΔW_{ICL}.

2.2 Dataset Pruning

Investigating the data redundant problem not only helps to improve the training efficiency but also helps us understand the representation ability of small data and how many training samples are required and sufficient for a learning system. (Yang et al., 2022) proposed to use the Influence Function to accurately and fast estimate the parameter change caused by weighting an example p for the training dataset. The influence of weighting p on the parameters is given by:

$$\mathcal{I}_{\mathrm{param}}(p) = \left.\frac{\mathrm{d}\hat{\theta}_{\delta,p}}{\mathrm{d}\delta}\right|_{\delta=0} = -H_{\hat{\theta}}^{-1}\nabla_\theta L(p, \hat{\theta}), \tag{2}$$

where $H_{\hat{\theta}} = \frac{1}{n}\sum_{p_i \in \mathcal{D}} \nabla^2_\theta L(p_i, \hat{\theta})$ is the Hessian and positive definite by assumption, $\mathcal{I}_{\mathrm{param}}(p) \in \mathbb{R}^N$, N is the number of network parameters, $\mathcal{D}$ is the original dataset. After getting the weighting of each example p, (Yang et al., 2022) propose generalization-guaranteed pruning or cardinality-guaranteed pruning to get the final compressed dataset $\hat{\mathcal{D}}$.

3 Method

3.1 Recall

Given the training dataset $\mathcal{D}$ and the query source x, we use BM25 (Robertson et al., 2009) to retrieve a set of relevant examples C_1 for x. For each example p_j in $\mathcal{D}$, the BM25 score, denoted as $R(p_j, x)$, is computed. This score reflects the relevance of example p_j to the query x. Specifically:

$$R_j = \mathrm{BM25}(p_j, x), \tag{3}$$

where R_j is the relevance score of example p_j with respect to query x.

Subsequently, we form the set C_1 which consists of the top-N examples with the highest relevance scores:

$$C_1 = \{p_j | j = 1, 2, \ldots, N\}, \tag{4}$$

where N is the number of examples we wish to recall for the given query x.

3.2 Influence-Awared Rerank

For each p in C_1, we calculate the input representation of tokens p as P and the meta-gradient $G_p = W_V P$. To compute $\mathcal{I}_{\mathrm{param}}(p)$ in Eq. (2), we require the Hessian of p for the parameter W_V, which necessitates the computation of second-order derivatives. However, we only have access to first-order derivatives approximations of the parameters. Considering that LLMs typically employ cross-entropy loss and maximum likelihood estimation (MLE) for fine-tuning, we have opted to employ the Fisher matrix as an approximation of the Hessian (Barshan et al., 2020). The key to the approximation process is as follows:

$$\nabla^2 f(\mathbf{x}) \approx \nabla f(\mathbf{x})\nabla f(\mathbf{x})^\top \tag{5}$$

Then, combining the Eq. (2) with Eq. (5), the expression of the influence function for p is:

$$\mathcal{I}_{\mathrm{param}}(p) = -\hat{H}_{\hat{\theta}}^{-1} G_p, \tag{6}$$

where $H_{\hat{\theta}} = \frac{1}{n}\sum_{p_i \in \mathcal{D}} G_p G_p^\top$.

The score of C_1 is determined by a combination of the influence score and the relevance score, represented as:

$$\mathcal{S} = \left\{\|\mathcal{I}_{\mathrm{param}}(p_1)\|_F^2 + R_1, \ldots, \|\mathcal{I}_{\mathrm{param}}(p_N)\|_F^2 + R_N\right\}.$$

Finally, the K in-context learning examples in C_2 are chosen by:

$$C_2 = \{p_i | i \in I\}\text{, where } I = \arg\max_{\substack{I \subseteq \{1,2,\ldots,|\mathcal{S}|\} \\ |I| = K}} \mathcal{S}.$$

$K = 3$		GPT2-XL		GPT2-Large		GPT2-Small		GPT2-Medium	
		Acc (%)	F_1 (%)	Acc (%)	F_1 (%)	Acc (%)	F_1 (%)	Acc (%)	F_1 (%)
Sick	BM25	42.63	33.01	27.68	26.72	31.72	23.34	31.52	26.45
	Ours	**47.07**	**35.28**	**31.11**	**28.51**	**35.35**	**26.25**	**32.53**	**27.16**
Cola	BM25	61.84	54.83	63.09	50.24	**65.96**	**48.79**	60.98	50.04
	Ours	**63.09**	**55.53**	**64.24**	**50.74**	65.87	48.39	**63.95**	**52.80**
Ethos-disability	BM25	77.01	57.44	82.76	62.42	68.97	56.92	74.71	50.26
	Ours	**83.91**	**66.17**	**87.36**	**64.14**	**74.71**	**62.96**	**77.01**	**51.67**
Tweet_eval_stance_feminist	BM25	**50.75**	**46.19**	44.78	40.96	41.79	31.64	44.78	**41.33**
	Ours	**50.75**	43.27	**46.27**	**41.88**	**43.28**	**32.01**	**46.27**	38.24
Tweet_eval_stance_hillary	BM25	49.28	40.63	42.03	41.12	42.03	41.12	46.38	**44.95**
	Ours	**53.62**	**40.68**	**53.62**	**51.45**	**46.38**	39.24	**50.72**	44.73
All dataset Avg	BM25	56.30	46.42	52.07	44.29	50.09	40.36	51.67	42.61
	Ours	**59.69**	**48.18**	**56.52**	**47.34**	**53.12**	**41.77**	**54.10**	**42.92**

Table 1: Results of four ICL examples. The boldface represents the best performance.

4 Experiments

In this section, we empirically verify the efficiency of our approach. The source code and all experiments have been shared at `https://anonymous.4open.science/r/ICL-F302`.

4.1 Experiments setup

This section introduces the detailed information about our experiments.

Models. We use the open source GPT2 model family (Radford et al., 2019) (i.e., GPT2-Small, GPT2-Medium, GPT2-Large, GPT2-XL) as a representative of large models to verify the effectiveness of our method.

Datasets. We use five datasets spanning four tasks: linguistic analysis, hate speech detection, tweet classification, and semantic similarity. Specifically, we employ *Linguistic Acceptability dataset (Cola)* (Warstadt et al., 2019), *online hate speech detection dataset (Ethos and Ethos-disability)* (Mollas et al., 2022), *Tweet_eval-stance_feminist* and *Tweet_eval_stance_hillary* (Barbieri et al., 2020) from Twitter, and *Sentences Involving Compositional Knowledge dataset (Sick)* (Marelli et al., 2014). We use Accuracy and F_1 score as evaluation metrics. Detailed dataset statistics and the prompt templates used can be found in Appendix A.2 and Appendix A.1.

Implementation Details. In the study, we chose $K = 3$ and $K = 4$ demonstrations to contrast example selection methods from training data. We set $N = 100$ for all models. Sentences were either truncated or supplemented to have a uniform length at 50% of the average sentence length. Although using multiple transformer layers' meta-gradient might be beneficial, considering the time efficiency, we used the first layer and obtained higher accuracy than baseline models.

Baselines. Considering the model proposed in this paper is unsupervised and requires no training, it possesses a higher generalizability and operational efficiency compared to models that undergo supervised training. To ensure a fair comparison, our primary baseline is the unsupervised BM25-based In-Context Example Selection. Previous work (Wang et al., 2023; Gupta et al., 2023) has demonstrated that BM25 constitutes a robust baseline for demonstration selection, hence we juxtapose our methodology against BM25. The demonstrations selected by the BM25 are utilized across all GPT2 models.

4.2 Overall Performance

Tables 1 and 2 display results for three and four ICL examples, respectively. Observing the last two rows, our method consistently outperforms across all models and datasets. Using three and four examples, we surpass BM25 by averages of 5.17% and 6.64% in all metrics. Specifically, accuracy sees improvements of 6.33% and 7.80% over BM25. This underscores our approach's superiority. We found

$K = 4$		GPT2-XL		GPT2-Large		GPT2-Small		GPT2-Medium	
		Acc (%)	F_1 (%)	Acc (%)	F_1 (%)	Acc (%)	F_1 (%)	Acc (%)	F_1 (%)
Sick	BM25	42.83	35.60	31.31	31.09	33.54	26.03	31.92	28.07
	Ours	**46.67**	**36.70**	**32.53**	**30.74**	**39.19**	**30.42**	**35.56**	**29.80**
Cola	BM25	60.98	**54.48**	62.80	51.17	65.10	48.26	60.21	49.85
	Ours	**61.36**	54.38	**63.47**	**51.51**	**67.31**	**48.71**	**64.91**	**54.05**
Ethos-disability	BM25	81.61	61.33	85.06	**61.60**	68.97	56.92	77.01	54.78
	Ours	**85.06**	**64.80**	**87.36**	59.87	**72.41**	**59.60**	**79.31**	**56.50**
Tweet_eval_stance_feminist	BM25	46.27	43.54	43.28	36.81	43.28	**40.64**	38.81	31.02
	Ours	**53.73**	**47.36**	**47.76**	**44.47**	**46.27**	35.10	**44.78**	**37.11**
Tweet_eval_stance_hillary	BM25	47.83	40.64	34.78	34.36	39.13	33.92	40.58	39.71
	Ours	**47.83**	**38.89**	**46.38**	**46.01**	**46.38**	**34.81**	**47.83**	**44.07**
All dataset Avg	BM25	55.90	47.12	51.45	43.01	50.00	41.16	49.71	40.69
	Ours	**58.93**	**48.42**	**55.50**	**46.52**	**54.31**	**41.73**	**54.48**	**44.30**

Table 2: Results of four ICL examples. The boldface represents the best performance.

some higher model performance with three ICL examples compared to four, which can be explained by overfitting and example quality. Overfitting in few-shot learning means too many examples leads to adaptation to specific instances rather than general patterns, reducing accuracy on unseen data. Furthermore, if the additional fourth example is of lower quality or relevance, it can degrade model performance.

5 Related Work

5.1 In-context Learning

In-context learning (ICL) has emerged as a fresh approach in natural language processing (NLP), where large models predict based solely on contexts supplemented by several examples (Dong et al., 2022; Shin et al., 2022; Zhang et al., 2023; Bai et al., 2023). Numerous studies have sought to modify, improve, and comprehend ICL, encompassing topics like prompt tuning (Kim et al., 2022; Wang et al., 2022a; Mishra et al., 2022), intrinsic mechanism analysis (Chan et al., 2022; Li et al., 2023; Garg et al., 2022), evaluations (Srivastava et al., 2023; Wang et al., 2022b), and its use across various fields (Sun, 2023), among others.

5.2 Demonstration Selection

The goal of demonstration selection is to identify optimal examples for ICL. (Liu et al., 2022) demonstrated that choosing the nearest neighbors as in-context examples is an effective approach. The used distance measures include the pre-set L2 distance or the cosine similarity based on sentence embeddings. They introduced KATE, an unsupervised kNN retriever for in-context example selection. (Rubin et al., 2022) suggested a two-phase retrieval process for demonstration selection. For a given input, it initially employs an unsupervised retriever (like BM25) to retrieve similar candidate examples and then uses a supervised retriever, EPR, to pick demonstrations from these candidates. Recent studies indicate that LLMs exhibit strong sensitivity to the examples chosen, resulting in significant performance variations (Nie et al., 2022), dependency on example sequence (Lu et al., 2022), and at times, an insensitivity to the actual labels (Min et al., 2022). Our research focuses on reducing training overhead and condensing crucial data from the training set into in-context examples, which in turn amplifies the ICL's effectiveness.

6 Conclusion

In summary, inspired by LLMs and ICL's potential, we devised a two-stage method using the influence function for optimal in-context example selection. Our approach ensures relevance with the query source and efficiently determines influence scores. The result is an enhancement in ICL performance, with our experiments validating our model's effectiveness. Our framework stands out due to its data-independent nature and minimal reliance on model parameters.

7 Limitation and Future Work

Given resource constraints and page limitations, we provide limited validation in this paper. Our model is a model-agnostic and free-training approach that can be applied to various in-context learning selection models. In the future, we will validate the effectiveness of our model on more large-scale language models, baselines, and datasets.

References

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. 2023. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *arXiv preprint arXiv:2306.04637*.

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.

Elnaz Barshan, Marc-Etienne Brunet, and Gintare Karolina Dziugaite. 2020. Relatif: Identifying explanatory training samples via relative influence. In *International Conference on Artificial Intelligence and Statistics*, pages 1899–1909. PMLR.

Stephanie Chan, Adam Santoro, Andrew Lampinen, Jane Wang, Aaditya Singh, Pierre Richemond, James McClelland, and Felix Hill. 2022. Data distributional properties drive emergent in-context learning in transformers. *Advances in Neural Information Processing Systems*, 35:18878–18891.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Zhifang Sui, and Furu Wei. 2022. Why can gpt learn in-context? language models secretly perform gradient descent as meta optimizers. *arXiv preprint arXiv:2212.10559*.

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 preprint arXiv:2301.00234*.

Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. 2022. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598.

Shivanshu Gupta, Sameer Singh, and Matt Gardner. 2023. Coverage-based example selection for in-context learning. *arXiv preprint arXiv:2305.14907*.

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 preprint arXiv:2206.08082*.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pages 1885–1894. PMLR.

Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. 2023. Transformers as algorithms: Generalization and stability in in-context learning. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pages 19565–19594. PMLR.

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.

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.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

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.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487.

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.

Feng Nie, Meixi Chen, Zhirui Zhang, and Xu Cheng. 2022. Improving few-shot performance of language models via nearest neighbor calibration. *arXiv preprint arXiv:2212.02216*.

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.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

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.

Seongjin Shin, Sang-Woo Lee, Hwijeen Ahn, Sungdong Kim, HyoungSeok Kim, Boseop Kim, Kyunghyun Cho, Gichang Lee, Woomyoung Park, Jung-Woo Ha, et al. 2022. On the effect of pretraining corpora on in-context learning by a large-scale language model. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5168–5186.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Zhongxiang Sun. 2023. A short survey of viewing large language models in legal aspect. *arXiv preprint arXiv:2303.09136*.

Boshi Wang, Xiang Deng, and Huan Sun. 2022a. Iteratively prompt pre-trained language models for chain of thought. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2714–2730.

Liang Wang, Nan Yang, and Furu Wei. 2023. Learning to retrieve in-context examples for large language models. *arXiv preprint arXiv:2307.07164*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022b. Super-naturalinstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *EMNLP*, pages 5085–5109. Association for Computational Linguistics.

Dataset	Training Size	Valid Size	Label class
Sick	4439	495	3
Cola	8551	1043	2
Ethos-disability	346	87	2
Tweet_eval-stance_feminist	597	67	3
Tweet_eval-stance_hillaryt	620	69	3

Table 3: Dataset statistics.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Shuo Yang, Zeke Xie, Hanyu Peng, Min Xu, Mingming Sun, and Ping Li. 2022. Dataset pruning: Reducing training data by examining generalization influence. *arXiv preprint arXiv:2205.09329*.

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. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2422–2437.

Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. 2023. Trained transformers learn linear models in-context. *arXiv preprint arXiv:2306.09927*.

A APPENDIX FOR REPRODUCIBILITY

A.1 Dataset

In our experiments, we employ *Linguistic Acceptability dataset (Cola)* (Warstadt et al., 2019), *online hate speech detection dataset (Ethos and Ethos-disability)* (Mollas et al., 2022), *Tweet_eval-stance_feminist* and *Tweet_eval_stance_hillary* (Barbieri et al., 2020) from Twitter, and *Sentences Involving Compositional Knowledge dataset (Sick)* (Marelli et al., 2014). Table 3 shows more detailed statics of the datasets.

A.2 Templates.

Based on (Min et al., 2022; Yoo et al., 2022), we employ minimal templates. For the GPT-2 series model, we distinguish between input and label using a space, and this is also applied between demonstration examples. Table 4 presents templates from five datasets, where blue represents the fixed section, and black varies depending on the example content.

Dataset	example	label
Sick	[S1] The question is:[S2] True or False? Answer:[Label]	True,Not sure,False
Cola	Review: [S1] Acceptability:[Label]	unacceptable,acceptable
Ethos-disability	There is a hate based on disability in [S1]True or False? Answer: [Label]	False,True
Tweet_eval-stance_feminist	Tweet: [S1]Sentiment: [Label]	none,against,favor
Tweet_eval-stance_hillary	Tweet: [S1]Sentiment: [Label]	none,against,favor

Table 4: Template examples.