Inspiring Out-of-distribution Generalization of In-Context Learning via Contrastive Demonstrations

Anonymous ACL submission

Abstract

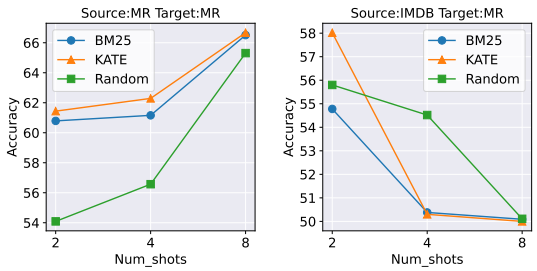
Although existing demonstration construction methods have significantly improved the performance of In-Context Learning (ICL), these unfortunately only focused on the in-distribution settings that the selected demonstrations should have the same distribution with testing data. However, the out-of-distribution (OOD) settings are more commonly encountered in real scenarios, but ignored in the age of large language models. This paper first investigates the performance of existing ICL demonstration construction methods in OOD settings and verifies their failures. Moreover, this paper proposes contrastive demonstrations that combine a demonstration with its counterfactual, where a rationale-guided counterfactual generation method is proposed to generate higher-quality counterfactual data. Extensive experiments validate the effectiveness of our proposed method and the contrastive demonstrations can help the model better identify the essence of the task, thus achieving OOD generalization.

1 Introduction

Large language models (LLMs) have shown in-context learning (ICL) capabilities with the increase in model size (Brown et al., 2020; Dong et al., 2022). Different from traditional paradigm, ICL enables the LLMs to adapt to the downstream tasks without parameter updating. Recent studies (Dong et al., 2022) have shown that with a few examples as prompts, LLMs could achieve even surpass the performance of full data fine-tuning.

To further boost the ICL performance across various tasks, recent studies focus on demonstration selection (Liu et al., 2022; Rubin et al., 2022; Yang et al., 2023), demonstration permutation (Lu et al., 2022), and demonstration format (Min et al., 2022).

However, almost all current studies focus on in-distribution performance but ignore out-of-distribution (OOD) settings, which is common in real scenarios (Hendrycks et al., 2020). In specific,



(a) In-distribution (b) Out-of-distribution

Figure 1: Performance comparisons of different demonstration selection methods: random selection, BM25 (Wang et al., 2023), and KATE (Liu et al., 2022). We select IMDB (Maas et al., 2011) and MR (PANG, 2005) as the source datasets and choose MR as the target dataset to show in-distribution and OOD performance.

the OOD setting in ICL refers to the test instances and the available selection instances belong to the same task but exhibit different distributions (Liu et al., 2021). This problem has been widely investigated in deep learning. However, the OOD in ICL was seldom addressed recently because LLMs are trained with huge data and believed to have strong generalization capabilities. In ICL, the provided demonstrations should guide the LLMs to understand the essence of the task rather than the overfitting on some specific datasets. As a result, existing demonstration construction methods were believed to inspire the OOD generalization ability of LLMs. However, the reality is harsh. Through experiments, we observe that the OOD is still a serious problem in ICL, even performing demonstration selection. As shown in Figure 1, we conduct experiments on sentiment classification tasks using retrieval-based demonstration selection methods, which have been shown to be the best demonstration selection methods. These methods get good performance in in-distribution settings but poor performance in OOD settings, even worse than random selection.

To achieve OOD generalization in deep learn-

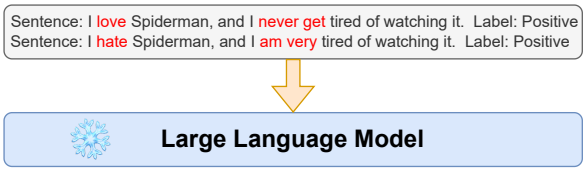


Figure 2: Example of contrastive demonstration.

ing, researchers usually adopt data augmentation, ensemble learning, and regularization (Liu et al., 2021). Among these methods, counterfactual augmented data is the simplest and the most effective method(Kaushik et al., 2019; Madaan et al., 2021). Compared to original data, such counterfactual augmented data makes minimal modifications to flip the label, which has been shown to effectively identify causally relevant features (Kaushik et al., 2020). Inspired by this, this paper proposes contrastive demonstrations, which are bound with original demonstrations as their counterfactuals. As shown in Figure 2, such contrastive demonstrations could effectively reveal the casual features of different labels. Following Wang et al. (2023), we compute the contribution of the text to the label and we find the casual features would contribute more in contrastive demonstrations, which indicates that such demonstrations could help the LLMs to know the essence of the task (Bhattacharjee et al., 2023). To construct contrastive demonstrations, we propose a rationale-guided counterfactual generation method. Firstly, Zhao et al. (2023) shows that LLM is a good rationale extractor. Inspired by this, the paper prompts the LLM to extract the rationale for the given instance. Then, we prompt the LLM to modify the rationale part to generate counterfactuals. Extensive experimental results show that such contrastive demonstrations could effectively inspire the OOD generalization ability of LLMs, bringing significant performance improvements in OOD settings. Our contributions can be summarized as follows:

- This paper first investigates the OOD setting in in-context learning and shows that existing demonstration construction methods are inadequate for solving this problem.
- To inspire the OOD generalization ability, this paper proposes contrastive demonstrations. Such demonstrations could better reveal the essence of the task and show significant performance improvements in OOD settings.

2 Preliminary

The success of counterfactual data augmentation in deep learning lies in the construction of counterfactual data that differs only in causal features from the original data while preserving similar non-causal features (Kaushik et al., 2020). As a result, models trained on such data naturally learn to capture these causal features, enabling generalization on OOD data. However, the model parameters are frozen under the ICL paradigm. Therefore, the effectiveness of counterfactual data remains underexplored.

Recently, Wang et al. (2023) shows the text information would aggregate into the label in the shallow layers. They define the contribution of the text to the label as follows:

$$I_l = \left| \sum_h A_{h,l} \odot \frac{\partial \mathcal{L}(x)}{\partial A_{h,l}} \right|. \quad (1)$$

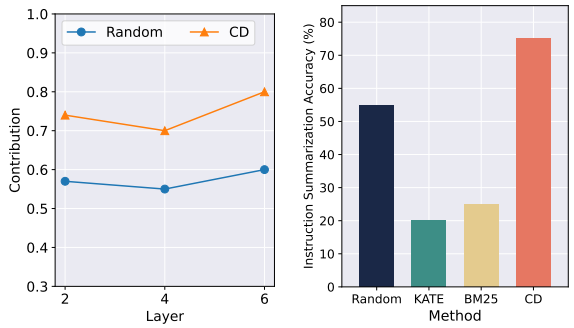
where $A_{h,l}$ is the value of the attention matrix of the h-th attention head in the l-th layer. $\mathcal{L}(x)$ is the loss function and x is the input. $I_l(i,j)$ represents the significance of the information flow from the j-th word to the i-th word for ICL. Following this definition, we compute the contribution of casual features and show the results in Figure 3(a). We observe that the casual feature in the contrastive demonstrations (CDs) would contribute more than others. Therefore, we believe the model also could learn the mapping from the casual feature to the label in contrastive demonstrations though their parameters remain unchanged, which has been explored from the view of implicit gradient descent (Von Oswald et al., 2023; Dai et al., 2023).

To further validate our conjecture, we conduct experiments on instruction summarization, which requires `gpt-3.5-turbo` to summarize the task instruction with the given input-output pairs. As shown in Figure 3(b), we find that contrastive demonstration (CD) achieves better summarization accuracy [1], which indicates that they can help the LLM better understand the essence of the task. Appendix B shows more details.

3 Method

We review the reasons for the effectiveness of counterfactual data, which is achieved by modifying causal features to flip labels while keeping non-causal features unchanged (Kaushik et al., 2020). Therefore, the key is to modify the casual features,

[1] 3 graduates annotate the summarization accuracy.



(a) Casual feature contribution in shallow layers for SST-2.

(b) Instruction summarization accuracy of for SST-2.

Figure 3: (a) Contribution (Wang et al., 2023) of the casual feature to the label for SST-2. (b) Instruction summarization accuracy of gpt-3.5-turo for SST-2.

also known as rationales. More recently, Zhao et al. (2023) shows LLM is a good rationale extractor. Inspired by this, we propose a rationale-guided counterfactual generation method.

Firstly, we prompt the LLM to find the rationale for the given sentence (Step 1). Then, with the identified rationale, we require the LLM to modify them to achieve label flipping (Step 2). The used templates are shown in Table 1. In this way, we can obtain the final counterfactual. In summary, we can obtain the counterfactual $e_i^{'} = (x_i^{'}, y_i^{'})$ for the original demonstration $e_i = (x_i, y_i)$.

For k-shot ICL, we usually select k demonstrations with different selection methods. As for contrastive demonstrations, we just need to select $k/2$ demonstrations and augment them with the counterfactuals to construct a k-shot demonstration:

$$\{e_1, e_2, \cdots, e_{K/2}\} \rightarrow \{e_1, e_1^{'}, e_2, e_2^{'} \cdots, e_{K/2}, e_{K/2}^{'}\} \tag{2}$$

In this paper, we obtain the original $k/2$ demonstrations by random sampling, which could better show the effectiveness of our method.

4 Experiments

4.1 Datasets

Following previous studies on OOD (Kaushik et al., 2019, 2020), we conduct experiments on the sentiment classification task. In specific, we utilize IMDB (Maas et al., 2011) as the source dataset and choose SST-2 (Socher et al., 2013) and MR (PANG, 2005) as the OOD testing dataset. More details about the datasets can be found in Appendix A.

Step 1 In the task of <task name>, the label of the following text is <y_i>, text: <x_i>. Explain why this text is <y_i> label by identifying the rationale, which refers to the words that caused the label.
Output <r_i>

Step 2 Generate the counterfactual of <x_i> with its rationale <r_i>, you can edit the rationale part of <x_i> to flip the original label <y_i>. And predict the label for the generated text.
Output <$x_i^{'}$> and <$y_i^{'}$>

Table 1: Rationale guided counterfactual generation.

4.2 Large Language Models

This paper conducts experiments on five large language models of different scales: GPT2-xl (1.5B) (Radford et al., 2019), GPT-J (6B) (Wang, 2021), LLaMA (7B) (Touvron et al., 2023a), LLaMA-2(7B) (Touvron et al., 2023b) and `gpt-3.5-turbo`.

4.3 Baselines

In this paper, we compare our method with the existing two types of demonstration selection methods for ICL as follows:

Instance-level Selection Methods usually retrieve demonstrations for each given test instance. We choose the retriever as semantic similarity (**KATE**) (Liu et al., 2022) and BM25 score (**BM25**) (Wang et al., 2022).

Task-level Selection Methods select demonstration for a provided dataset rather than specific instances, which could significantly improve efficiency (Yang et al., 2023). Following previous studies, we select demonstrations from different semantic clusters (**Cluster**) (Zhang et al., 2022; Gao et al., 2023) and select diverse demonstrations by determinantal point process (**DPP**) (Ye et al., 2023; Yang et al., 2023).

4.4 Main Results

Figure 4 shows the corresponding results. Both instance-level methods and task-level methods get poor performance in OOD settings, even worse than random sampling. Compared to these baselines, our methods significantly improve the performances across different LLMs. Besides, find that the baseline methods exhibit worse performance when k increases, which is in stark contrast to the in-domain results.

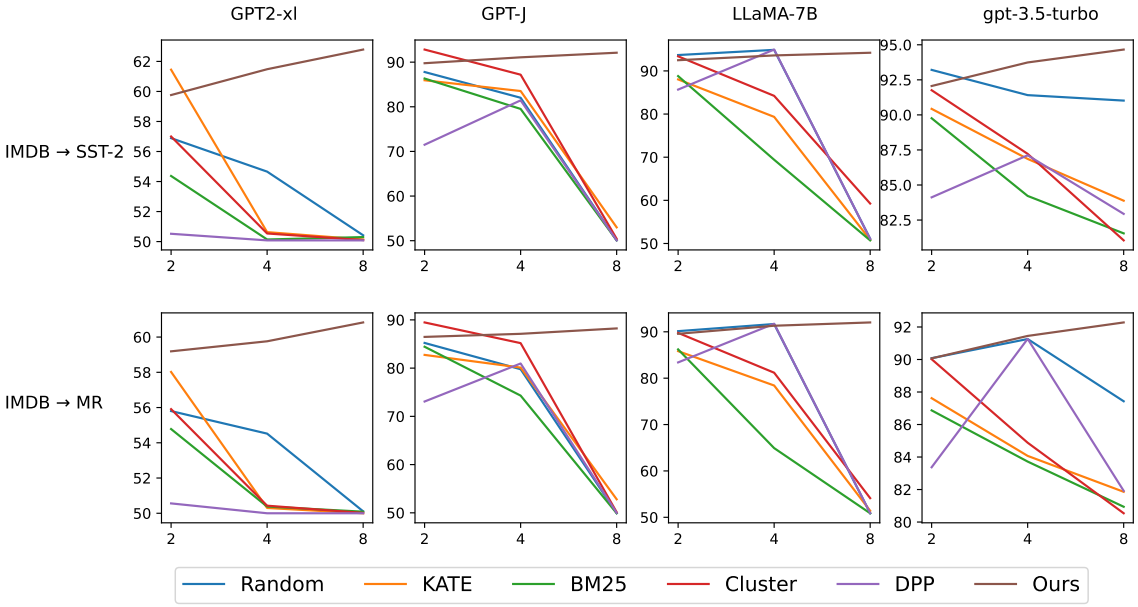


Figure 4: Main results of OOD setting. All results are reported with the mean of five runs.

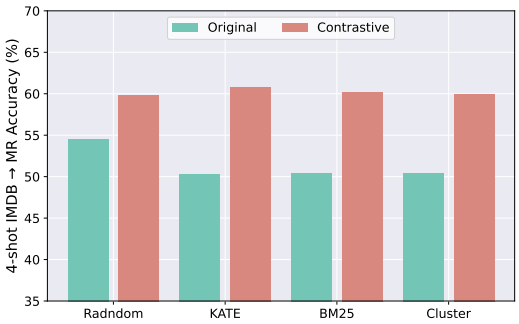


Figure 5: Effevctiveness on different selection methods.

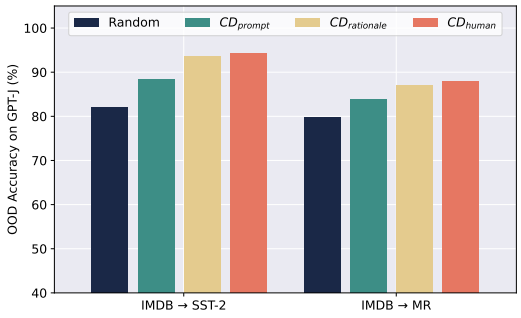


Figure 6: 4-shot accuracy of different variants.

4.5 Effect on Different Selection Methods

In the main experiments, we obtain the original demonstrations by random sampling. To show the effectiveness of contrastive demonstrations, we conduct experiments on different selection methods. Figure 5 presents the corresponding results. From these results, we find that contrastive demonstration could bring significant improvements in different demonstration selection methods.

4.6 Comparisons of Different Counterfactual Generation Methods

The core of contrastive demonstrations is the counterfactual construction, we consider different counterfactual construction methods in this section.

Our method utilizes the rationale and therefore we name it as $\text{CD}_{\texttt{rationale}}$. For naive prompting to generate counterfactual, we name this variant as $\text{CD}_{\texttt{prompt}}$. Kaushik et al. (2019) annotates counterfactuals by humans and we can also use these high-quality data, this variant is defined as $\text{CD}_{\texttt{human}}$. We compare these three variants.

Figure 6 presents the results on GPT-J. We observe that all variants achieve better performance. Besides, our proposed rationale-guided counterfactual generation method brings significant improvements to prompting methods and gets comparable performance with human-annotated data.

5 Conclusion

In this paper, we first investigate the OOD setting in ICL and show existing demonstration construction methods get poor performance in this setting. To solve this problem, we propose contrastive demonstrations and such demonstrations could reveal the essence of the task, inspiring the OOD generalization ability. Extensive experiments also validate the effectiveness of the proposed method.

Limitations

The main limitation of this paper is the proposed contrastive demonstration is built on counterfactual. However, the generation of counterfactuals poses a challenging problem for many complex tasks. Therefore, previous studies on counterfactual data augmentation have also focused on the simpler tasks that are examined in this paper. Additionally, the comparative nature of counterfactuals for complex tasks is not obvious enough, raising doubts about the effectiveness of contrastive demonstrations for complex tasks in the ICL setting. Hence, we consider the application of contrastive examples to complex tasks as our future work. The main contribution of this paper is the first investigation of the OOD setting in the ICL paradigm, which is an important but overlooked problem. And we propose contrastive demonstration and validate its effectiveness, the counterfactual generation is just a simple step in our method.

Ethics Statement

This paper investigates the OOD setting for in-context learning, and the experiments are conducted on publicly available datasets with available LLMs. As a result, there is no data privacy concern. Meanwhile, this paper does not involve human annotations, and there are no related ethical concerns.

References

Amrita Bhattacharjee, Raha Moraffah, Joshua Garland, and Huan Liu. 2023. Llms as counterfactual explanation modules: Can chatgpt explain black-box text classifiers? *arXiv preprint arXiv:2309.13340*.

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.

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.

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*.

Shuzheng Gao, Xin-Cheng Wen, Cuiyun Gao, Wenxuan Wang, and Michael R Lyu. 2023. Constructing effective in-context demonstration for code intelligence tasks: An empirical study. *arXiv preprint arXiv:2304.07575*.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020. Pretrained transformers improve out-of-distribution robustness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2744–2751.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

Divyansh Kaushik, Amrith Setlur, Eduard H Hovy, and Zachary Chase Lipton. 2020. Explaining the efficacy of counterfactually augmented data. In *International Conference on Learning Representations*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, William B 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.

Jiashuo Liu, Zheyan Shen, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. 2021. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*.

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.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Nishtha Madaan, Inkit Padhi, Naveen Panwar, and Diptikalyan Saha. 2021. Generate your counterfactuals: Towards controlled counterfactual generation for text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13516–13524.

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.

B PANG. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43nd Meeting of the Association for Computational Linguistics (ACL), 2005*.

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.

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.

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.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. 2023. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174. PMLR.

Ben Wang. 2021. Mesh-Transformer-JAX: Model-Parallel Implementation of Transformer Language Model with JAX. https://github.com/kingoflolz/mesh-transformer-jax.

Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023. Label words are anchors: An information flow perspective for understanding in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9840–9855, Singapore. Association for Computational Linguistics.

Shuohang Wang, Yichong Xu, Yuwei Fang, Yang Liu, Siqi Sun, Ruochen Xu, Chenguang Zhu, and Michael Zeng. 2022. Training data is more valuable than you think: A simple and effective method by retrieving from training data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3170–3179.

Zhao Yang, Yuanzhe Zhang, Dianbo Sui, Cao Liu, Jun Zhao, and Kang Liu. 2023. Representative demonstration selection for in-context learning with two-stage determinantal point process. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5443–5456.

Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. 2023. Compositional exemplars for in-context learning. *arXiv preprint arXiv:2302.05698*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2023. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*.

A Details of Datasets

Table 2 shows the data statistics of our used datasets in the experiments.

Dataset	Class	Train	Test
IMDB	2	25000	25000
SST-2	2	6921	1821
MR	2	8530	1066

Table 2: Stastics of the datasets.

The prompt format for these datasets is shown in Table 3.

B Details of Instruction Summarization

In the instruction summarization experiments, we require `gpt-3.5-turo` to generate the task instructions for the provided input-output pairs. Table 4 presents the prompt of this experiment.

We provide 100 4-shot input-out pairs and invite 3 graduates to annotate the correctness of the generated instructions. For random, we just random sample 4 instances from the training dataset. For KATE, we first randomly select one instance and retrieve 4 instances by semantic similarity. For BM25, we first randomly select one instance and retrieve 4 instances by semantic similarity. For CD, we first randomly select 2 instances and augment them with their counterfactuals to construct 4-shot input-output pairs.

Table 5 presents some cases of the generated instructions. To achieve OOD generalization, we hope the instruction is independent of any domain, such as the `foreign film` and `comedy` in the bad case.

C Details of Experiments

We illustrate the details of experiments in this section due to the space limitation.

C.1 Details of Main Experiments

Due to space limitations, we show the whole results in Table 6.

C.2 Experimental Details of Section 4.5

To construct contrastive demonstrations, we need to obtain the original $k/2$ demonstrations for k-shot ICL. In our main experiments, we get these $k/2$ by random selection. We explore the effect of different selection methods in this section. And we conduct experiments on IMDB $\rightarrow$ MR and report the results of 4-shot accuracy on GPT2-xl.

Prompt Format	Label Names
Sentence: `[sentence]` Sentiment:	Positive Negative

Table 3: Prompt formats of the three datasets.

I give a friend an instruction and some input. The friend reads the instruction and writes an output for every input. Here are the input-output pairs:
Input: x_1 Output: y_1
Input: x_2 Output: y_2
Input: x_3 Output: y_3
Input: x_4 Output: y_4
The instruction is:

Table 4: Prompt of the task instruction summarization experiments.

C.3 Experimental Details of Section 4.6

We compare different counterfactual generation methods in this section. According to this part, we can find the importance of the formulation of contrastive demonstrations.

Good Case

For each input, output whether the sentiment is positive or negative.
Provide the sentiment of the input.

Bad Case

Watch this foreign film and write your overall impression of it.
Read the review of comedy and determine if it is positive or negative.

Table 5: Example of generated instruction.

Model	Method	IMDB → SST-2	IMDB → MR
		2/4/8	2/4/8
GPT2-xl 1.5B	Random	56.89/54.66/50.42	55.80/54.52/50.11
	KATE	61.45/50.63/50.14	58.02/50.30/50.00
	BM25	54.37/50.14/50.30	54.78/50.38/50.09
	Cluster	57.00/50.54/50.10	55.91/50.43/50.02
	DPP	50.52/50.08/50.08	50.56/50.00/50.00
	CD	59.76/61.48/62.79	59.19/59.76/60.83
GPT-J 6B	Random	87.79/82.02/50.07	85.22/79.77/49.94
	KATE	85.94/83.53/52.99	82.74/80.11/52.81
	BM25	86.33/79.52/50.08	84.43/74.30/49.91
	Cluster	92.82/87.19/50.43	89.47/85.18/50.13
	DPP	71.50/81.44/50.08	73.08/80.96/50.00
	CD	89.76/91.08/92.11	86.47/87.11/88.23
LLaMA 7B	Random	93.68/94.88/51.08	90.14/91.68/50.88
	KATE	88.03/79.35/50.85	85.83/78.42/51.41
	BM25	88.80/69.41/50.72	86.21/64.92/50.87
	Cluster	93.38/84.22/59.26	89.77/81.18/54.13
	DPP	85.67/94.95/51.08	83.40/91.74/50.91
	CD	92.48/93.59/94.21	89.57/91.31/92.02
`gpt-3.5-turbo`	Random	93.21/91.41/91.02	90.08/91.26/87.43
	KATE	90.43/86.86/83.88	87.62/84.07/81.86
	BM25	89.76/84.22/81.55	86.87/83.72/80.94
	Cluster	91.76/87.22/81.05	90.02/84.88/80.54
	DPP	84.12/87.12/82.94	83.37/91.27/81.92
	CD	92.06/93.74/94.66	90.07/91.45/92.28

Table 6: 2-shot/4-shot/8-shot performance comparison.