ER-ICL: Error Book Maybe More Valuable for In-context Learning

Anonymous ACL submission

Abstract

In-context learning (ICL) with few-shot examples has emerged as a key strength of large language models (LLMs), allowing them to adapt to new tasks with just a few examples. Recent research suggests that ICL closely resembles implicit fine-tuning. Building on this, we hypothesize that demonstrations where LLMs make mistakes could offer stronger learning signals for ICL, potentially leading to enhanced performance compared to instances where LLMs predict correctly. To explore this, we created an 'Error Book' comprising such demonstrations, and used a retriever to select relevant instances from this collection instead of the entire training dataset. Our experiments across two different tasks show that this Error Book based Retrieval In-Context Learning (ER-ICL) not only boosts performance but also improves retrieval efficiency by reducing the search scope. Our results indicate that leveraging error-driven demonstrations could be a valuable strategy for enhancing in-context learning.

1 Introduction

In-context learning uses some examples to prompt the model while these examples are generally from training data or data from other similar tasks. This allows LLMs to be generalised to unfamiliar tasks without fine-tuning (Brown et al., 2020). There are lots of research try to understand why ICL work (Dong et al., 2022). A compelling theory suggests that ICL closely resembles implicit fine-tuning (Dai et al., 2022). The authors demonstrated, through theoretical and empirical analysis, that the Transformer attention mechanism (Vaswani et al., 2017) is akin to a form of gradient descent, which parallels the performance of ICL with explicit fine-tuning. In this work, we build upon this insight, seeking to enhance ICL performance.

To clarify, let's delve into the underlying concept. For fine-tuning to be effective, training data must provide loss values. This allows the model to adjust its incorrect predictions through back propagation. If ICL is approximated to fine-tuning, the in-context demonstrations should ideally be those where the model initially makes mistakes. Without such cases, there is no opportunity for gradient descent during ICL. Our hypothesis suggests that selecting demonstrations where the model is likely to make incorrect predictions will offer stronger learning signals during ICL, leading to potential improved performance compared to demonstrations where the model already predicts correctly.

To test this hypothesis, we construct a demonstration corpus, titled the 'Error Book', comprising only demonstrations where the model's predictions are inaccurate. This approach aims to leverage these mis-predictions as a potent tool for enhancing learning within the ICL framework. To obtain the 'Error Book', given a training dataset, we use zero-shot chain-of-thought prompting (Brown et al., 2020) and get the model's prediction of each data point. If the model predicts wrongly, we add it to the 'Error Book'. Our empirical results indicate that using randomly selected demonstrations from the 'Error Book' leads to improved performance compared to using a broader range of demonstrations from the entire training dataset.

Furthermore, recent advancements in Retrieval-based ICL have showcased that choosing demonstrations similar to the input query leads to superior performance, compared to random or manually curated selections. Building on this, we introduce a two-stage pipeline (illustrated in Figure 1) to further improve the ICL performance of LLMs. The initial stage involves constructing the 'Error Book' as previously described. In the subsequent stage, we employ a retrieval system to select k demonstrations for use in ICL.

By conducting experiments on two tasks with a range of in-context demonstrations and two retrieval techniques (BM25 (Robertson et al., 2009) and Top-K semantic selections (Arora et al., 2017)),

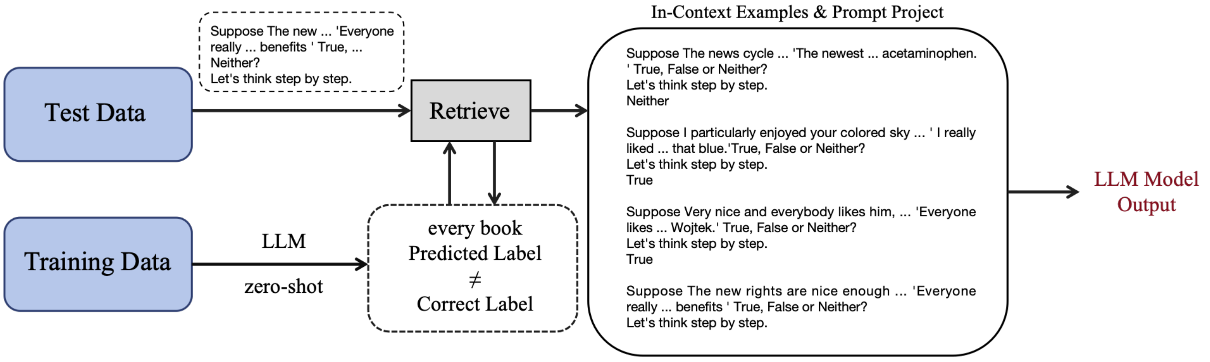

Figure 1: Overview of our proposed method for selecting ICL demonstrations: 1) Using LLMs to identify samples with model prediction errors (Error Book). 2) Retrieving k samples as demonstrations and selecting prompt and predicting results.

we showcase that the proposed method outperforms the baseline of retrieving from the entire training set. We observed an average improvement of 2.77% on the RTE (Wang et al., 2019) and 0.85% on the Winograd (ai2, 2019) dataset. Moreover, this approach not only enhances performance but also reduces inference time, owing to a smaller retrieval space. These findings support previous hypothesis about the approximation of In-Context Learning (ICL) to fine-tuning and our work introduce a new, more efficient ICL framework.

2 Related work

How to select the retrieval corpus Due to the fact that the demonstrations of in-context learning are all obtained from the retrieval corpus, the composition of the retrieval corpus greatly affects the performance of in-context learning. Z-ILC (Lyu et al., 2022) produces pseudo samples and their corresponding pseudo labels via LLM to as retrieval corpus, subsequently diminishing the overhead associated with manual sample and label generation. Gao et al. (2023) use an off-the-shelf retriever to retrieve samples of LLM prediction errors that fall within an ambiguous label set as demonstrations. UDR (Li et al., 2023) serves as a universal retriever across diverse tasks. By add different prompts into the existing retrieval corpus, creating a new retrieval corpus for different tasks. Our work is based on a comprehension of why ICL functions effectively. This understanding has guided us to develop a corpus consisting of demonstrations where LLMs tend to make mistakes.

Retrieval augmented In-context learning for LLM Retrieval Augmented LLMs combine LLM generative abilities with retrieval techniques. Vote-k (Su et al., 2022) and CEIL (Ye et al., 2023) enhance in-context learning by selecting contextually relevant exemplars and modeling interrelationships among them. Other approaches, like (Dalvi et al., 2022), incorporate supplementary explanations during retrieval. DSP (Khattab et al., 2022) and MOT (Li and Qiu, 2023) present frameworks for iterative retrieval and categorization of demonstrations. LLM-R (Rubin et al., 2021), UPRISE (Cheng et al., 2023), and Dr.ICL (Luo et al., 2023) focus on demonstration categorization for retriever retraining. Lastly, RetICL (Scarlatos and Lan, 2023) and (Lu et al., 2022) apply reinforcement learning in in-context learning for LLM retrieval model training. In our study, we showcase that the effectiveness of retrieval-augmented ICL can be further improved by taking selective demonstrations for retrieval.

3 Methodology

Ours methodology unfolds in two main phases:

Step 1: Construct Error Book Specifically, Let $D_{train} = \{(x_0, y_0), (x_1, y_1), ..., (x_n, y_n)\}$ denotes our training data. We first employ template to convert x_i into a question-answering format using a template (see Table 6). Then, we execute CoT zero-shot prompting to derive the answer to the question. An LLM predicts the outcome $\hat{Y} = \{\hat{y}(x_0), \hat{y}(x_1), ..., \hat{y}(x_n)\}$. Subsequently, we select the incorrectly predicted instances $\hat{D}_{train}$ from the training samples to form the 'Error Book' as the retrieval corpus,

$$\hat{D}_{train} = \{x_i \in D | I(\hat{y}(x_i) \neq y_i) = 1\} .$$

where I denote identification function. Selection of instances exhibiting model prediction errors within the training data.

Step2: Retrieval of Similar Demonstrations We employed two popular retrieval methods, BM25 (Robertson et al., 2009) and Top-k (Liu et al., 2021). BM25 emphasizes term matching while Top-k focuses on semantic matching. Top-k first uses a certain sentence encoder to sentences $\{x_1, x_1, ..., x_k\}$ in both the training set and test set to vector representations $\{v_1, v_2, ..., v_k\}$. This method retrieve the k nearest neighbors to each test sample from the training set as demonstrations, utilizing the vector encoded by the sentence encoder. These methods, representing distinct retrieval mechanisms, complement each other and showcase our method's effectiveness across different retrieval styles.

4 Experiments

Large Language Model We experiment with the large language model LLaMA-2 (7B) (Touvron et al., 2023) based on pre-training and fine-tuning. Our framework is not limited to LLaMA but work for other models, and we leave the exploration as a future work.

Tasks We conducted experiments using two distinct datasets. Recognizing Textual Entailment (RTE) datasets corresponds to one of the NLI tasks featured in SuperGLUE (Wang et al., 2019) with labels: neutral, entailment and contradiction. The other one is Winograd (ai2, 2019) which is a commonsense reasoning task that involves selecting the correct option for a given sentence, with the options represented by specific, concrete words.

Baseline and Experiment Setting The baseline approach we employ involves retrieving from the entire training data (termed as entire book). To eliminate errors in the results arising from variations in the number of demonstrations, each retrieve selects demonstration samples in quantities of 1, 5, 10, and 15 for experimental.

5 Experiment Results

5.1 Random Selection Pipeline

First of all, we demonstrate the effectiveness of 'Error Book' even not in the retrieval setting. Here, after the 'Error Book' is constructed, we randomly select the demonstrations from it and compare with the baseline that randomly selected from the entire dataset. The results are presented in Table 1. In the RET dataset, utilizing the Error Book as the retrieval corpus led to an average improvement of 3.5% compared to entire book. In the Winograd dataset, results from employing the Error Book surpassed those of the entire book in two of the four settings, but were relatively inferior in the other two. We hypothesize that this variation is due to the heavy reliance of Winograd performance on commonsense knowledge. If the underlying knowledge in the Error Book differs from that in the input query, the demonstrations are less likely to be significantly helpful.

Datastes	Type	1	5	10	15
RTE	1	56.3	70.0	70.7	71.4
	2	**56.6**	**74.7**	**75.4**	**75.8**
Winograd	1	28.5	**52.5**	52.8	**54.1**
	2	**31.0**	52.1	**52.9**	53.6

Table 1: Comparison of using 'Entire Book' (Type1) and 'Error Book' (Type2) for randomly selecting demonstrations.

5.2 Retrieval Pipeline

RTE In Figure 2 (the top two subfigures), we show that utilizing both BM25 and Top-k, retrieving from 'Error Book' consistently surpassed the baseline. The performance improvements were quantified as an average increase of 1.2% for BM25, and 3.4% for Top-k retrievers. We also observed that with the gradual increase in the number of demonstrations, the performance enhancement of our method compared to the baseline becomes progressively more significant.

Winograd Our analysis reveals divergent outcomes when employing Top-k and BM25 methods. The 'Error Book' demonstrates a more pronounced improvement with the Top-k retriever, enhancing the baseline by an average of 1.5%. Conversely, the BM25 retriever's performance aligns more closely with random selection, with the 'Error Book' not offering significant benefits, and even underperforming in some cases. We suggest that semantic-based retrievers like Top-K are more effective than BM25 or random selection, especially for commonsense reasoning tasks. Therefore, combining a robust retriever such as Top-K with the 'Error Book' emerges as the most effective approach for enhancing commonsense reasoning.

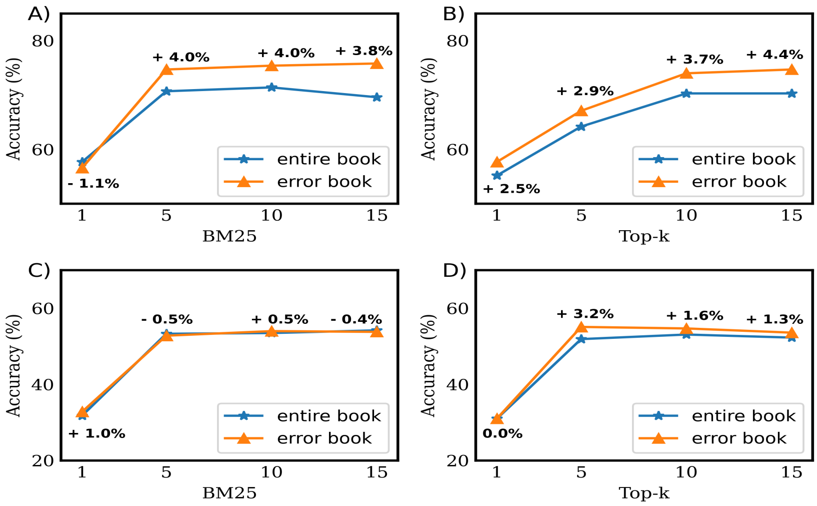

Figure 2: Comparison of our method and baseline under different number of demonstration scenarios using different retrievers on different datasets. A): using BM25 on the RET dataset; B): Using Top-k on the RET dataset; C): Using BM25 on the Winograd dataset; D): Using Top-k on the Winograd dataset.

dataset	'Error Book' percentage
RTE	48
Winograd	58

Table 2: Percentage of the Error Book in different datasets

Setting	Random	BM25	Top-k
1	**75.4**	**73.2**	**74.0**
2	70.7	70.3	68.5
3	66.7	66.0	66.0

Table 3: Setting 1: retrieve from 'Error Book'; Setting 2: retrieve from 'Error Book' and 'Correct Book'; Setting 3: retrieve from 'Correct Book'.

To summarize, our method has yielded improvement compared to the baseline over different number of shots as in-context learning demonstrations. The experiment also revealed that since the size corpus of the Error Book is only half that of the entire book (see Table 2), the retrieval time when using the Error Book is consequently reduced to half of that required for using the entire book.

5.3 Ablation Study

To test our hypothesis that better results correlate with more learning signals in demonstrations, we created a 'Correct Book' corpus which is the exclusive set to the 'Error Book' and designed three experimental settings with varying learning signals. The first setting exclusively used 'Error Book' demonstrations for maximum learning signal. The second mixed 'Correct Book' and 'Error Book' demonstrations for a moderate signal. The third used only 'Correct Book' demonstrations, offering the least signal. Each experiment used 10 demonstrations. The results, detailed in Table 3, show a consistent decline across settings, supporting our hypothesis that stronger learning signals in demonstrations lead to better results.

6 Conclusion

In our study, we introduce an innovative method that leverages 'Error Books' for in-context learning demonstration retrieval. This idea is predicated on the concept that in-context learning is similar to the gradient descent process seen in fine-tuning. We argue that demonstrations where the model initially predicts incorrectly offer more learning potential and are therefore more effective than those where the model predicts correctly. To validate this argument, we conducted experiments across two different tasks and three varied settings for obtaining demonstrations (retrieval-based and non-retrieval based). The results clearly show that our method surpasses the baseline in performance, underscoring its effectiveness. These results also lead us to an intriguing new line of inquiry: exploring how 'Error Books' used as a retrieval corpus can improve a model's contextual learning and adaptation. This exploration not only fuels our curiosity but could also shed light on the intricacies of error correction and learning in Large Language Models.

Limitation

In our research, we successfully showcase the efficacy of our methods on two tasks and with one Large Language Model (LLM). While incorporating additional datasets and models could strengthen our evidence, this initial exploration into the error correction capabilities of LLMs has revealed promising results and interesting trends, setting a solid foundation for further study in this area.

References

2019. Winogrande: An adversarial winograd schema challenge at scale.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *International conference on learning representations*.

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

Daixuan Cheng, Shaohan Huang, Junyu Bi, Yuefeng Zhan, Jianfeng Liu, Yujing Wang, Hao Sun, Furu Wei, Denvy Deng, and Qi Zhang. 2023. Uprise: Universal prompt retrieval for improving zero-shot evaluation. *arXiv preprint arXiv:2303.08518*.

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*.

Bhavana Dalvi, Oyvind Tafjord, and Peter Clark. 2022. Towards teachable reasoning systems. *arXiv preprint arXiv:2204.13074*.

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*.

Lingyu Gao, Aditi Chaudhary, Krishna Srinivasan, Kazuma Hashimoto, Karthik Raman, and Michael Bendersky. 2023. Ambiguity-aware in-context learning with large language models. *arXiv preprint arXiv:2309.07900*.

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp. *arXiv preprint arXiv:2212.14024*.

Xiaonan Li, Kai Lv, Hang Yan, Tianyang Lin, Wei Zhu, Yuan Ni, Guotong Xie, Xiaoling Wang, and Xipeng Qiu. 2023. Unified demonstration retriever for in-context learning. *arXiv preprint arXiv:2305.04320*.

Xiaonan Li and Xipeng Qiu. 2023. Mot: Pre-thinking and recalling enable chatgpt to self-improve with memory-of-thoughts. *arXiv preprint arXiv:2305.05181*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2022. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *arXiv preprint arXiv:2209.14610*.

Man Luo, Xin Xu, Zhuyun Dai, Panupong Pasupat, Mehran Kazemi, Chitta Baral, Vaiva Imbrasaite, and Vincent Y Zhao. 2023. Dr. icl: Demonstration-retrieved in-context learning. *arXiv preprint arXiv:2305.14128*.

Xinxi Lyu, Sewon Min, Iz Beltagy, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. Z-icl: Zero-shot in-context learning with pseudo-demonstrations. *arXiv preprint arXiv:2212.09865*.

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. 2021. Learning to retrieve prompts for in-context learning. *arXiv preprint arXiv:2112.08633*.

Alexander Scarlatos and Andrew Lan. 2023. Reticl: Sequential retrieval of in-context examples with reinforcement learning. *arXiv preprint arXiv:2305.14502*.

Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, et al. 2022. Selective annotation makes language models better few-shot learners. *arXiv preprint arXiv:2209.01975*.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint arXiv:1905.00537*.

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

A Experiment Codebase.

We use Openicl, an open-source framework for in-context learning. https://github.com/Shark-NLP/OpenICL.

B Quantitative analysis

We present some examples in Table 4, where retrieving from the 'Error Book' is better than retrieving from the entire training data. Despite the apparent semantic correlation of the demonstration selected using the 'Entire Book' being higher with the given example, it was observed that the prediction accuracy was actually superior in the case of the example selected using the 'Error Book'. This outcome suggests a nuanced interaction between semantic relevance and prediction accuracy in our model.

C Our method and baseline comparison

The comparison of accuracy between our model and baseline on different datasets and retrievers is shown in Table 7

dataset	entire book	Error Book	input query and truth answer
RTE	In 2005, Leland Yee, a ... Can we infer California Assembly Bills 1792 & 1793 are laws against ultraviolent video games.? True or False? Let's think step by step. True	In an interview on a television... Can we infer Scientology is against psychiatry.? True or False? Let's think step by step. True	A U.S. Court of Appeals on ... Can we infer California Assembly Bills 1792 & 1793 are laws against ultraviolent video games.? True or False? Let's think step by step. True
Winograd	Sentence: The business had to... drawers or desks? Let's think step by step. drawers	Sentence: The household was preparing to send the children to ... pencils or pens? Let's think step by step. pens	Sentence: The students were at... desks or pencils? Let's think step by step. desks

Table 4: Demonstration retrieved on entire book and Error Book using the BM25 retriever. You can select just the top-1 example.

dataset	entire book	Error Book	input query
Winograd	Sentence: The girl put **bread or waffle?**... Sentence: James needed... **eggplant or pot?** eggplant Sentence: She used... **pasta or eggplant?** pasta Sentence: She used... **pasta or eggplant?** eggplant Sentence: I used... **toaster or oven?** toaster	Sentence: I tried to... **pipe or brush?**... brush Sentence: It took... **pumpkin or eggplant?**... eggplant Sentence: Mary put the... **pan or oven?**... pan Sentence: The girl... **bread or waffle?**... waffle Sentence: She used **pasta or eggplant?**... eggplant	Sentence: Terry tried... **eggplant or toaster?**...

Table 5: the demonstration of using BM25 in entire book and Error Book.

dataset	template
RTE	</text1> Can we infer </text2>? True or False? Let's think step by step. Answer:
Winograd	</sentence> Replace the _ in the above sentence with the correct option: - </option1> - </option2> Let's think step by step. Answer:

Table 6: The templates for each dataset.

Dataset	Method	Retrival	1	5	10	15
RTE	Baselines	Random	56.3	70.0	70.7	71.4
		BM25	57.7	70.7	71.4	69.6
		Top-k	55.2	64.2	70.3	70.3
	ours	Random	56.6	74.7	75.4	75.8
		BM25	56.3	71.4	73.2	73.6
		Top-k	57.7	67.1	74.0	74.7
Winograd	Baselines	Random	28.5	52.5	52.8	54.1
		BM25	31.8	53.3	53.5	54.2
		Top-k	31.0	51.9	53.1	52.3
	ours	Random	31.0	52.1	52.9	53.6
		BM25	32.8	52.8	54.0	53.8
		Top-k	31.0	55.1	54.7	53.6

Table 7: The comparison of accuracy between our model and baseline.