TUTOR-ICL: Guiding Large Language Models for Improved In-Context Learning Performance

Anonymous ACL submission

Abstract

There has been a growing body of work focusing on the in-context learning (ICL) abilities of large language models (LLMs). However, it is an open question of how effective ICL can be. This paper presents TUTOR-ICL, a simple prompting method that guides LLMs through the ICL process, inspired by how effective instructors might engage their students in learning a task. Specifically, we propose presenting exemplar answers in a comparative format rather than the traditional single-answer format. We also show that including the test instance before the exemplars can improve performance, making it easier for LLMs to focus on relevant exemplars. Lastly, we include a summarization step before attempting the test, following a common human practice. Experiments on various classification tasks, conducted across both decoder-only LLMs (Llama 2, 3) and encoder-decoder LLMs (Flan-T5-XL, XXL), show that TUTOR-ICL consistently boosts performance, achieving up to a 13.76% increase in accuracy.

1 Introduction

With the rapid advancement of large language models (LLMs), in-context learning (ICL), which involves performing various tasks by learning from only a small number of examples within the context of a single prompt, has become a dominant paradigm in natural language processing (Brown et al., 2020). With ICL, the likelihood of any answer for the test example is conditioned on the provided ICL exemplars (Dong et al., 2022). The underlying assumption of ICL is that LLMs can thoroughly review these exemplars, identify the hidden patterns crucial for input-label mappings, and consequently make correct predictions (Wang et al., 2023). However, recent studies have provided some evidence that LLMs, particularly smaller ones (less than or around 10 billion parameters), are unable to fully utilize the provided exemplars. For example, Shivagunde et al. (2024) found that smaller LLMs allocate less attention mass to ICL examples than larger models. Similarly, Wei et al. (2023) indicated that smaller LLMs have a lower ability to adjust their semantic priors based on the provided ICL examples than their larger counterparts. We also conduct a preliminary experiment indicating that LLMs do not always produce the correct answer, even when it is provided as one of the exemplars (Section 5.1 and Table 3). These results strongly suggest that a lot of LLMs still struggle in performing ICL.

Our Objective and Approach In this paper, we address this limitation by investigating the following research question: *How can we effectively guide LLMs to achieve better ICL performance?* Our solution is to enhance the prompt template with simple yet powerful ideas inspired by how intelligent humans would perform ICL: (1) framing ICL as a comparative reading task; (2) showing the test example early to make it easier to identify and focus on relevant exemplars; (3) summarizing the material before the test to organize and digest the learned knowledge.

We first introduce the concept of a **comparative answer** format. In contrast to most prior works that offer a single answer (e.g., “positive”), we suggest presenting the answer in a comparative format (e.g., “closer to positive than neutral”). This straightforward adjustment results in a notable performance boost, such as an average F1 increase of 5.78 points on the Laptop14 ABSC dataset with Llama3-8B-Instruct. Additionally, inspired by how people generally start by reading new information, then take a moment to digest and summarize the newfound knowledge before proceeding, we incorporate a **“summarization” step** into the prompt. Lastly, we present the **glance-at-the-test (GAT)** framework which is driven by the idea that knowing the test example in advance could encourage a more efficient search and concentration on the exemplars relevant to the test. We show that incorporating these new elements into a single ICL prompt improves performance of a number of LLMs (Llama2-7B,13B, Llama3-8B-Instruct, Flan-T5-XL,XXL) on a variety of text classification tasks (aspect-based sentiment classification, news topic classification, and question type classification).

2 Related Work

General studies on ICL The existing literature on in-context learning research can be broadly divided into two categories: (1) analytical studies, which aim to uncover the underlying mechanisms of how LLMs perform ICL (Wang et al., 2023; Yoo et al., 2022; Von Oswald et al., 2023), and (2) improvement studies, which seek to enhance ICL performance through various methods such as exemplar selection (Liu et al., 2021; Ye et al., 2023), exemplar ordering (Min et al., 2022), and instruction calibration (Zhou et al., 2022). Relatively few studies have focused on the prompt *template* itself, which is the specific area where our work lies, as explained next.

Studies on ICL template components A number of studies have examined the effects of the components within prompt templates. Shivagunde et al. (2024), referred to as Decon-ICL hereafter, showed the benefits of briefly repeating text and the importance of reiterating inline instructions. Xu et al. (2023) proposed RE2 as a simple strategy of reading the question again to improve ICL performance. Wei et al. (2023) experimented the effects of flipped or semantically-unrelated labels on ICL.

Most of these studies rely on a standardized prompt template that includes four components: task instructions, exemplar inputs, exemplar labels, and inline instructions (Shivagunde et al., 2024). The key distinction of our work lies in our out-of-the-box approach. Rather than focusing solely on these four standard components, we take it one step further by exploring the incorporation of new elements. We verify that these additional elements can significantly improve ICL performance.

Efficient LLM decoding Another prominent line of research focuses on intermediate step reasoning. Recently, various prompting techniques have emerged, including chain-of-thought (Wei et al., 2022), self-consistency (Wang et al., 2022), tree-of-thought (Yao et al., 2024), and visualization-of-thought (Wu et al., 2024), among others. However, these methods share common limitations, such as excessive computational costs due to sampling-based prompting and the challenge of obtaining high-quality exemplars for use in ICL. Therefore, in our study we focus on the standard greedy decoding setting which requires just a single API-call.

3 TUTOR-ICL

3.1 Motivation

Our work is primarily motivated by Shivagunde et al. (2024) and Xu et al. (2023), which show that simply repeating the same instruction or question could boost LLM performance. This indicates that providing more careful guidance could further enhance LLMs' ICL performance, highlighting a potential area for improvement.

Given this motivation, we introduce TUTOR-ICL, which effectively guides LLMs throughout the ICL process by incorporating three novel ideas: (1) Comparative answer format, which provides the answers in a comparative form to elicit deeper thinking from multiple answer perspectives; (2) Glance-at-the-test framework, which informs LLMs of the test instance in advance, leading to a more efficient search and focus on relevant exemplars; (3) Summarization step, which makes LLMs to summarize the given exemplars before attempting the test instance, similar to human practice.

3.2 Comparative Answer Format

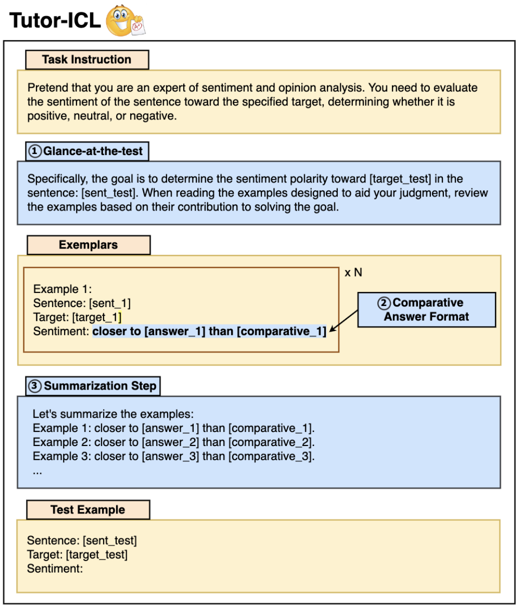
As in Figure 1-②, we provide answers in a comparative format (e.g., "closer to positive than neutral") rather than the traditional single-answer format (e.g., "positive"). The rationale behind this approach is that LLMs would produce answers in a comparative format by following the exemplars. This automatically leads them to compare different answers, thereby encouraging deeper thinking from various answer perspectives. More details on selecting the comparative answer and phrasing the overall answer are described in Appendix C.

3.3 Glance-at-the-Test Framework

The majority of ICL studies present the test instance at the end. However, our investigation reveals that presenting the test instance at the beginning, as well as at the end, is often beneficial. Intuitively, when the test instance is given in advance, LLMs can leverage this prior information to concentrate more on the relevant exemplars, by using their self-attention mechanism (Vaswani et al., 2017). This is not possible for decoder-only LLMs if the test instance is presented only at the end. An example is provided in Figure 1-①.

3.4 Summarizing before the Test

Summarizing is a vital skill for humans to organize and gain a deeper understanding of material. We examine whether this is also true for LLMs. After reading the exemplars, we add a brief summary of them before solving the test instance as illustrated in Figure 1-③. There could be many ways to create a summary, but for the sake of simplicity, we chose repeating the answers as the default approach.



Pretend that you are an expert of sentiment and opinion analysis. You need to evaluate the sentiment of the sentence toward the specified target, determining whether it is positive, neutral, or negative.

Specifically, the goal is to determine the sentiment polarity toward [target_test] in the sentence: [sent_test]. When reading the examples designed to aid your judgment, review the examples based on their contribution to solving the goal.

Example 1:
Sentence: [sent_1]
Target: [target_1]
Sentiment: **closer to [answer_1] than [comparative_1]**

Let's summarize the examples:
Example 1: closer to [answer_1] than [comparative_1].
Example 2: closer to [answer_2] than [comparative_2].
Example 3: closer to [answer_3] than [comparative_3].
...

Sentence: [sent_test]
Target: [target_test]
Sentiment:

Figure 1: The overall template of TUTOR-ICL. The three main components of TUTOR-ICL are represented in blue. Best viewed in color.

4 Experiments

4.1 Experimental Settings

Datasets We selected three widely used classifications tasks in ICL: aspect-based sentiment classification (ABSC) (SemEval-14-Laptops and Restaurants (Pontiki et al., 2014)), news topic classification (AGNews (Zhang et al., 2015)), and question type classification (TREC QC (Li and Roth, 2002)). Detailed explanations for each task can be found in Appendix A. We form the validation set by collecting 300 instances for each label from the training set. The ICL exemplars are randomly selected for each seed from the remaining training data, and the final evaluation is conducted on the test set.

Models and Settings We utilize both encoder-decoder LLMs (Flan-T5-XL,XXL) (Chung et al., 2024) and decoder-only LLMs (Llama2-7B,13B, and Llama3-8B-Instruct) (Touvron et al., 2023). We use n exemplars for each answer label: $n = 1$ for AGNews and TREC QC, and $n = 2$ for ABSC. More details can be found in Appendix B.

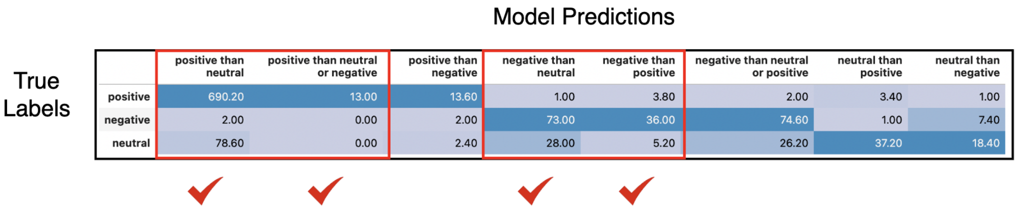
4.2 Results

Main Results Tables 1 and 2 present the performance of TUTOR-ICL and the baseline methods on the test set. We can see that TUTOR-ICL consistently enhances performance across all models and datasets, showing the greatest improvement in TREC QC with Llama3-8B-Instruct, where the accuracy increases by 13.76% and F1 score by 12.70 points. Additionally, TUTOR-ICL surpasses relevant competitors, such as RE2 (Xu et al., 2023) and Decon-ICL (Shivagunde et al., 2024) styles.

Ablation Results We provide detailed ablation results in Table 5 in the Appendix. We chose the best-performing models from each category as representatives: Flan-T5-XXL for encoder-decoder LLMs and Llama3-8B-Instruct for decoder-only LLMs. We observe that each method is generally effective on its own, and combining them results in even further improvement.

	Rest14		Lap14	
Model	Acc (%)	F1 (%)	Acc (%)	F1 (%)
1. Flan-T5-XL (3B)				
• Baseline-ICL	82.73$_{0.57}$	68.65$_{2.09}$	77.96$_{0.35}$	71.61$_{0.87}$
• RE2-style	82.68$_{0.48}$	67.70$_{0.38}$	77.32$_{0.40}$	70.88$_{0.95}$
• Decon-ICL-style	81.48$_{0.41}$	64.80$_{1.13}$	77.52$_{0.42}$	69.96$_{0.96}$
• TUTOR-ICL	**83.89**$_{\mathbf{0.45}}$	**72.02**$_{\mathbf{1.32}}$	**80.72**$_{\mathbf{0.64}}$	**76.64**$_{\mathbf{1.04}}$
2. Flan-T5-XXL (11B)				
• Baseline-ICL	84.87$_{0.19}$	71.61$_{0.69}$	81.53$_{0.30}$	75.08$_{0.51}$
• RE2-style	84.00$_{0.48}$	69.15$_{1.38}$	80.00$_{0.61}$	72.32$_{1.05}$
• Decon-ICL-style	83.18$_{0.36}$	66.46$_{1.00}$	79.47$_{0.19}$	71.16$_{0.25}$
• TUTOR-ICL	**87.43**$_{\mathbf{0.23}}$	**79.21**$_{\mathbf{0.55}}$	**84.92**$_{\mathbf{0.53}}$	**81.13**$_{\mathbf{0.66}}$
3. Llama2 (7B)				
• Baseline-ICL	66.29$_{3.17}$	55.71$_{2.84}$	57.77$_{1.76}$	52.27$_{1.91}$
• RE2-style	60.12$_{2.96}$	55.05$_{2.62}$	55.38$_{2.17}$	53.59$_{2.02}$
• Decon-ICL-style	64.46$_{2.98}$	55.84$_{2.85}$	57.21$_{3.01}$	52.57$_{2.71}$
• TUTOR-ICL	**71.91**$_{\mathbf{3.18}}$	**60.37**$_{\mathbf{3.58}}$	**63.13**$_{\mathbf{1.74}}$	**58.57**$_{\mathbf{2.19}}$
4. Llama2 (13B)				
• Baseline-ICL	78.48$_{1.16}$	67.59$_{1.76}$	73.41$_{1.07}$	65.49$_{2.06}$
• RE2-style	77.55$_{1.48}$	65.51$_{2.12}$	72.89$_{1.20}$	64.56$_{1.99}$
• Decon-ICL-style	80.21$_{1.13}$	68.44$_{1.88}$	74.40$_{1.18}$	66.80$_{1.52}$
• TUTOR-ICL	**82.83**$_{\mathbf{0.91}}$	**71.82**$_{\mathbf{2.29}}$	**77.40**$_{\mathbf{0.89}}$	**71.90**$_{\mathbf{1.01}}$
5. Llama3-8B-Instruct				
• Baseline-ICL	83.00$_{0.25}$	67.37$_{1.11}$	76.40$_{0.71}$	66.36$_{1.86}$
• RE2-style	82.79$_{0.67}$	66.46$_{2.40}$	76.02$_{0.93}$	65.58$_{2.21}$
• Decon-ICL-style	83.20$_{0.97}$	67.21$_{3.01}$	76.55$_{0.24}$	66.50$_{0.86}$
• TUTOR-ICL	**84.55**$_{\mathbf{0.46}}$	**75.32**$_{\mathbf{1.23}}$	**81.47**$_{\mathbf{0.59}}$	**77.32**$_{\mathbf{0.97}}$

Table 1: Few-shot ICL results on ABSC. Average of five random seeds and standard deviations in the subscript.



Model Predictions

True Labels	positive than neutral	positive than neutral or negative	positive than negative	negative than neutral	negative than positive	negative than neutral or positive	neutral than positive	neutral than negative
positive	690.20	13.00	13.60	1.00	3.80	2.00	3.40	1.00
negative	2.00	0.00	2.00	73.00	36.00	74.60	1.00	7.40
neutral	78.60	0.00	2.40	28.00	5.20	26.20	37.20	18.40
	✓	✓		✓	✓			

Figure 2: Further evidence that the comparative answer format actually triggers comparative reasoning in LLMs. New types of comparative answers (indicated by red checkmarks) are frequently generated.

	AGNews		TREC	
Model	Acc (%)	F1 (%)	Acc (%)	F1 (%)
1. Flan-T5-XXL (11B)				
• Baseline-ICL	$92.09_{0.05}$	$92.09_{0.05}$	$93.44_{0.43}$	$91.81_{0.38}$
• RE2-style	$91.60_{0.07}$	$91.61_{0.07}$	$94.04_{0.32}$	$92.46_{0.29}$
• Decon-ICL-style	$91.65_{0.21}$	$91.65_{0.22}$	$93.52_{0.30}$	$91.84_{0.27}$
• TUTOR-ICL	$\mathbf{92.30}_{\mathbf{0.06}}$	$\mathbf{92.28}_{\mathbf{0.06}}$	$\mathbf{95.00}_{\mathbf{0.18}}$	$\mathbf{93.49}_{\mathbf{0.47}}$
2. Llama3-8B-Instruct				
• Baseline-ICL	$79.62_{2.51}$	$78.78_{3.43}$	$63.40_{2.38}$	$63.09_{2.39}$
• RE2-style	$79.53_{2.32}$	$78.40_{3.01}$	$63.88_{2.50}$	$64.34_{2.50}$
• Decon-ICL-style	$81.02_{2.01}$	$80.36_{2.52}$	$66.72_{1.60}$	$64.02_{1.67}$
• TUTOR-ICL	$\mathbf{83.42}_{\mathbf{1.92}}$	$\mathbf{83.13}_{\mathbf{2.34}}$	$\mathbf{77.16}_{\mathbf{1.08}}$	$\mathbf{75.79}_{\mathbf{1.96}}$

Table 2: Overall results on AGNews and TREC.

	Rest14		Lap14	
Model	Acc (%)	F1 (%)	Acc (%)	F1 (%)
1. Llama3-8B-Instruct				
• Baseline-ICL (gold@first)	$99.50_{0.16}$	$99.11_{0.27}$	$98.84_{0.44}$	$98.61_{0.59}$
• TUTOR-ICL (gold@first)	$\mathbf{99.96}_{\mathbf{0.08}}$	$\mathbf{99.94}_{\mathbf{0.13}}$	$\mathbf{99.78}_{\mathbf{0.18}}$	$\mathbf{99.76}_{\mathbf{0.19}}$
• Baseline-ICL (gold@last)	$96.89_{0.38}$	$94.81_{0.70}$	$94.76_{2.15}$	$93.72_{2.67}$
• TUTOR-ICL (gold@last)	$\mathbf{98.77}_{\mathbf{0.34}}$	$\mathbf{98.07}_{\mathbf{0.48}}$	$\mathbf{96.61}_{\mathbf{0.59}}$	$\mathbf{96.10}_{\mathbf{0.77}}$
2. Flan-T5-XXL (11B)				
• Baseline-ICL (gold@first)	$97.27_{0.14}$	$95.31_{0.25}$	$95.45_{0.46}$	$93.92_{0.60}$
• TUTOR-ICL (gold@first)	$\mathbf{97.86}_{\mathbf{0.22}}$	$\mathbf{96.45}_{\mathbf{0.34}}$	$\mathbf{95.83}_{\mathbf{0.33}}$	$\mathbf{94.47}_{\mathbf{0.43}}$
• Baseline-ICL (gold@last)	$97.84_{0.12}$	$96.35_{0.23}$	$96.27_{0.17}$	$94.94_{0.22}$
• TUTOR-ICL (gold@last)	$\mathbf{98.77}_{\mathbf{0.04}}$	$\mathbf{98.00}_{\mathbf{0.08}}$	$\mathbf{97.58}_{\mathbf{0.50}}$	$\mathbf{96.79}_{\mathbf{0.60}}$

Table 3: TUTOR-ICL triggers deeper examination.

5 Analysis

5.1 Does TUTOR-ICL really help LLMs to more thoroughly examine the exemplars?

Beyond the performance improvement, we seek additional evidence to verify whether TUTOR-ICL is truly encouraging LLMs to more thoroughly examine the ICL exemplars. To this end, we designed a straightforward experiment as follows. The idea is to include the test instance (test sentence and answer) as one of the exemplars. Intuitively, if the LLM reads the exemplars thoroughly, accuracy should approach 100%, since the answer is given. We compare the baseline template with the TUTOR-ICL template in two scenarios (test instance as the first or last exemplar) as shown in Table 3. The results indicate that the TUTOR-ICL template consistently achieves higher accuracy, suggesting it enables LLMs to examine the exemplars more thoroughly.

5.2 Does comparative answer really trigger comparative thinking in LLMs?

Beyond the performance improvement, we offer deeper insights into the effectiveness of the comparative answer format. We design an experiment to verify whether this format genuinely triggers comparative reasoning in LLMs. Our hypothesis is that *"If the LLM generates a comparative answer not presented in the exemplars, it indicates that LLMs are not merely copying the labels but are actually engaging in comparative reasoning on their own."* To this end, we conduct an experiment to investigate whether the model can generate novel types of comparative answers that were not included in the exemplars. Specifically, we only provide "closer to positive than negative", "closer to negative than neutral or positive", "closer to neutral than positive", and "closer to neutral than negative" in the exemplars. As illustrated in Figure 2, we observe that four new answer types are generated from Flan-T5-XXL on the Rest14 dataset, averaged over 5 seeds. This verifies that the comparative answer format can indeed stimulate comparative reasoning rather than simply replicating the provided labels.

6 Conclusion

This paper has proposed an original framework, TUTOR-ICL, integrating three novel concepts into the standard in-context learning (ICL) prompt template: the comparative answer format, the glance-at-the-test framework, and the summarization step. To the best of our knowledge, TUTOR-ICL is the first work to incorporate new components to the ICL template, highlighting a new potential direction for future developments in the field.

7 Limitations

Our work has several limitations. Firstly, our study focused exclusively on classification tasks and did not extend to generative tasks. Additionally, due to hardware limitations, our analysis primarily involved models with up to 13 billion parameters. Exploring the effectiveness of TUTOR-ICL on significantly larger models would be an interesting future work. Lastly, as discussed in Section 2, our focus on greedy decoding was driven by computational efficiency. Nevertheless, investigating the integration of TUTOR-ICL with sampling-based prompting techniques remains a promising area for further exploration.

References

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.

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

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

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*.

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? *arXiv preprint arXiv:2202.12837*.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Namrata Shivagunde, Vladislav Lialin, Sherin Muckatira, and Anna Rumshisky. 2024. Deconstructing in-context learning: Understanding prompts via corruption. *arXiv preprint arXiv:2404.02054*.

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*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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.

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. *arXiv preprint arXiv:2305.14160*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. 2023. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*.

Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. 2024. Visualization-of-thought elicits spatial reasoning in large language models. *arXiv preprint arXiv:2404.03622*.

Xiaohan Xu, Chongyang Tao, Tao Shen, Can Xu, Hongbo Xu, Guodong Long, and Jian-guang Lou. 2023. Re-reading improves reasoning in language models. *arXiv preprint arXiv:2309.06275*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. 2023. Compositional exemplars for in-context learning. In *International Conference on Machine Learning*, pages 39818–39833. PMLR.

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. *arXiv preprint arXiv:2205.12685*.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *NIPS*.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*.

A Tasks and Datasets

In our study, we employ three classification tasks: aspect-based sentiment classification (ABSC), topic classification, and question type classification. Specifically, we use SemEval-14-Laptops and Restaurants datasets for ABSC (Pontiki et al., 2014), Ag's news topic classification dataset for topic classification (Zhang et al., 2015), Text REtrieval Conference Question Classification (TREC QC) dataset for question type classification (Li and Roth, 2002).

Laptops and Restaurants are collections of laptop and restaurant reviews where the task is to evaluate the sentiment (positive, neutral, or negative) of the review toward a specified target within the sentence. We selected this ABSC task for sentiment classification since it is a more challenging variant.

AGNews is a task to classify the given news article into one of the four categories: world, sports, business, or sci/tech.

TREC QC is a task to classify the given question into one of the six categories: abbreviation, entity, description, human, location, or number.

Detailed statistics for each dataset are provided in Table 4. All experiments are conducted on a single NVIDIA A100 GPU.

	Dataset		Label Words	
Task	Train	Test	Label	Count
Lap14	2313	638	Positive	341
			Negative	128
			Neutral	169
Rest14	3602	1120	Positive	728
			Negative	196
			Neutral	196
AGNews	120000	7600	World	1900
			Sports	1900
			Business	1900
			Sci/Fi	1900
TREC QC	5452	500	Abbreviation	9
			Entity	94
			Description	138
			Human	65
			Location	81
			Number	113

Table 4: Detailed information on the sizes of the training and test datasets for each task, as well as the sizes of the test datasets for each label within each task.

B Choosing baseline prompts and number of exemplars

Since the prompt might be sensitive to sentence phrasing, we experimented with five paraphrased instructions generated by ChatGPT[1] and selected the one with the best validation performance as the baseline. The five specific paraphrases are listed below. Option five, which was generally effective across most models, was chosen as the baseline instruction. Similar for AGNews and TREC QC.

1. Pretend that you are an expert of sentiment and opinion analysis. For a given sentence and a target, you have to assess the sentiment polarity (positive, neutral, or negative) towards the target.

2. Pretend that you are an expert of sentiment and opinion analysis. For a given sentence and a target, you have to assess the sentiment of the sentence toward the target, determining whether it is positive, neutral, or negative.

3. Pretend that you are an expert of sentiment and opinion analysis. Given a sentence and a target, you need to determine the sentiment of the sentence toward the target as either positive, neutral, or negative.

4. Pretend that you are an expert of sentiment and opinion analysis. For the provided sentence and target, your task is to assess the sentiment toward the target, identifying it as positive, neutral, or negative.

5. Pretend that you are an expert of sentiment and opinion analysis. You need to evaluate the sentiment of the sentence toward the specified target, determining whether it is positive, neutral, or negative.

[1] ChatGPT, March, 2024, OpenAI, https://chat.openai.com.

Number of exemplars used. We use n exemplars for each answer label: $n = 1$ for AGNews and TREC QC, and $n = 2$ for ABSC. We experimented with $n = 1, 2,$ and 3. For ABSC, $n = 2$ yielded the best baseline performance. For AGNews and TREC QC, both $n = 1$ and $n = 2$ showed similar results, so we selected $n = 1$ considering the inference speed.

C TUTOR-ICL Prompt Templates

C.1 Selecting Comparative Answers

To select the comparative answer corresponding to an answer we follow the below simple rules:

ABSC

- For positive label, we use neutral as the default comparative answer.
- For negative label, we use neutral as the default comparative answer.
- For neutral label, we use both positive and negative as default comparative answers.

AGNews and TREC We simply choose the next label based on the instruction as the comparative answer.

C.2 TUTOR-ICL Template Examples

Examples of TUTOR-ICL templates are provided below. We use $n = 2$ for ABSC and $n = 1$ for AGNews and TREC QC as described in B. Each prompt comprises five components: task instruction, Glance-at-the-Test (GAT) framework, exemplars with comparative answers, summary, and test. Minor adjustments are made based on the validation performance.

TUTOR-ICL for ABSC

Pretend that you are an expert of sentiment and opinion analysis. You need to evaluate the sentiment of the sentence toward the specified target, determining whether it is positive, neutral, or negative.

Specifically, the goal is to determine the sentiment polarity toward [target_test] in the sentence: [sent_test]. When reading the examples designed to aid your judgment, review the examples based on their contribution to solving the goal.

Example 1:
Sentence: [sent_1]
Target: [target_1]
Answer: closer to [answer_1] than [comparative_1].

Example 2:
Sentence: [sent_2]
Target: [target_2]
Answer: closer to [answer_2] than [comparative_2].

Example 3:
Sentence: [sent_3]
Target: [target_3]
Answer: closer to [answer_3] than [comparative_3].

Example 4:
Sentence: [sent_4]
Target: [target_4]
Answer: closer to [answer_4] than [comparative_4].

Example 5:
Sentence: [sent_5]
Target: [target_5]
Answer: closer to [answer_5] than [comparative_5].

Example 6:
Sentence: [sent_6]
Target: [target_6]
Answer: closer to [answer_6] than [comparative_6].

Let's summarize the examples:
example 1: closer to [answer_1] than [comparative_1].
example 2: closer to [answer_2] than [comparative_2].
example 3: closer to [answer_3] than [comparative_3].
example 4: closer to [answer_4] than [comparative_4].
example 5: closer to [answer_5] than [comparative_5].
example 6: closer to [answer_6] than [comparative_6].

Now use the above examples to solve your goal. When you find an answer, verify the answer carefully by comparing with the provided examples. Include verifiable evidence in your reasoning.

Sentence: [sent_test]
Target: [target_test]
Answer:

	Rest14		Lap14		AGNews		TREC	
Model	Acc (%)	F1(%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)
Flan-T5-XXL (11B)								
1. Baseline	$84.87_{0.19}$	$71.61_{0.69}$	$81.53_{0.30}$	$75.08_{0.51}$	$92.09_{0.05}$	$92.09_{0.05}$	$93.44_{0.43}$	$91.81_{0.38}$
2.1 Baseline + Comp.Ans.	$86.41_{0.23}$	$75.60_{0.63}$	$83.51_{0.57}$	$78.43_{1.02}$	$92.15_{0.08}$	$92.15_{0.08}$	$93.72_{0.20}$	$92.06_{0.22}$
2.2 Baseline + GAT	$86.38_{0.27}$	$75.88_{0.60}$	$82.13_{0.46}$	$76.43_{0.63}$	$92.18_{0.07}$	$92.17_{0.08}$	$94.20_{0.25}$	$93.19_{0.19}$
2.3 Baseline + Summary	$85.25_{0.39}$	$72.36_{1.22}$	$81.97_{0.60}$	$75.90_{1.03}$	$92.24_{0.12}$	$92.23_{0.12}$	$94.20_{0.13}$	$92.57_{0.11}$
3. TUTOR-ICL	$\mathbf{87.43_{0.23}}$	$\mathbf{79.21_{0.55}}$	$\mathbf{84.92_{0.53}}$	$\mathbf{81.13_{0.66}}$	$\mathbf{92.30_{0.06}}$	$\mathbf{92.28_{0.06}}$	$\mathbf{95.00_{0.18}}$	$\mathbf{93.49_{0.47}}$
Llama3-8B-Instruct								
1. Baseline	$83.00_{0.25}$	$67.37_{1.11}$	$76.40_{0.71}$	$66.36_{1.86}$	$79.62_{2.51}$	$78.78_{3.43}$	$63.40_{2.38}$	$63.09_{2.39}$
2.1 Baseline + Comp.Ans.	$84.04_{0.41}$	$72.42_{1.12}$	$77.99_{0.58}$	$72.14_{1.14}$	$81.85_{2.70}$	$81.18_{3.49}$	$64.80_{2.41}$	$63.76_{1.64}$
2.2 Baseline + GAT	$84.11_{0.62}$	$72.72_{1.54}$	$78.84_{0.82}$	$72.31_{1.31}$	$80.59_{2.22}$	$79.89_{2.75}$	$75.92_{0.73}$	$75.50_{1.08}$
2.3 Baseline + Summary	$83.41_{0.71}$	$69.54_{2.17}$	$77.46_{0.36}$	$69.87_{1.04}$	$80.74_{2.68}$	$80.09_{3.51}$	$70.76_{1.50}$	$70.11_{1.53}$
3. TUTOR-ICL	$\mathbf{84.55_{0.46}}$	$\mathbf{75.32_{1.23}}$	$\mathbf{81.47_{0.59}}$	$\mathbf{77.32_{0.97}}$	$\mathbf{83.42_{1.92}}$	$\mathbf{83.13_{2.34}}$	$\mathbf{77.16_{1.08}}$	$\mathbf{75.79_{1.96}}$

Table 5: Ablation study results using few-shot ICL. Comp.Ans.: Comparative Answer, GAT: Glance-at-the-Test framework. We chose the best performing model from each category (i.e., Llama3 for decoder-only and Flan-T5-XXL for encoder-decoder LLMs).

TUTOR-ICL for AGNews

Pretend that you are an expert in topic classification. For a given news article, you need to assess the topic of the article, determining whether it is world, sports, business, or sci/tech.

Specifically, the goal is to determine the topic of the news: [sent_test]. When reading the examples designed to aid your judgment, review the examples based on their contribution to solving the goal.

Example 1:
News: [sent_1]
Answer: The topic is closer to [answer_1] than [comparative_1].

Example 2:
News: [sent_2]
Answer: The topic is closer to [answer_2] than [comparative_2].

Example 3:
News: [sent_3]
Answer: The topic is closer to [answer_3] than [comparative_3].

Example 4:
News: [sent_4]
Answer: The topic is closer to [answer_4] than [comparative_4].

Let's summarize the examples so far:
example 1: [sent_1] | [answer_1].
example 2: [sent_2] | [answer_2].
example 3: [sent_3] | [answer_3].
example 4: [sent_4] | [answer_4].

Now use the above examples to solve your goal. When you find an answer, verify the answer carefully by comparing with the provided examples. Include verifiable evidence in your reasoning.

News: [sent_test]
Answer: The topic is closer to

Tutor-ICL for TREC QC

Pretend that you are an expert in question classification. You need to classify the question into one of the following semantic classes: abbreviations, entities, description, humans, location, or numerical.

Specifically, the goal is to determine the semantic class of the question: [sent_test]. When reading the examples designed to aid your judgment, review the examples based on their contribution to solving the goal.

Example 1:
Question: [sent_1]
Answer: Rather than [comparative_1], more accurately described as [answer_1].

Example 2:
Question: [sent_2]
Answer: Rather than [comparative_2], more accurately described as [answer_2].

Example 3:
Question: [sent_3]
Answer: Rather than [comparative_3], more accurately described as [answer_3].

Example 4:
Question: [sent_4]
Answer: Rather than [comparative_4], more accurately described as [answer_4].

Example 5:
Question: [sent_5]
Answer: Rather than [comparative_5], more accurately described as [answer_5].

Example 6:
Question: [sent_6]
Answer: Rather than [comparative_6], more accurately described as [answer_6].

Let's summarize the examples:
example 1: [sent_1] | Rather than [comparative_1], more accurately described as [answer_1].
example 2: [sent_2] | Rather than [comparative_2], more accurately described as [answer_2].
example 3: [sent_3] | Rather than [comparative_3], more accurately described as [answer_3].
example 4: [sent_4] | Rather than [comparative_4], more accurately described as [answer_4].
example 5: [sent_5] | Rather than [comparative_5], more accurately described as [answer_5].
example 6: [sent_6] | Rather than [comparative_6], more accurately described as [answer_6].

Now use the above examples to solve your goal. When you find an answer, verify the answer carefully by comparing it with the provided examples. Include verifiable evidence in your reasoning.

Question: [sent_test]
Answer: