In-Context Learning May Be Underestimated in Medical Domains

Anonymous ACL submission

Abstract

In medical domains, hospitals and medical research institutions produce large-scale real-world data with physician-annotated diagnoses every day. An ideal solution is to conduct fine-tuning (FT) with these data when developing large language models (LLMs) for medical domains. However, considering patients' privacy, it is still suspicious that de-identification is not performed carefully and LLMs may memorize the patient's information during FT. Instead, in-context learning (ICL) only relies on few-shot demonstrations. LLMs with ICL perform quite better than zero-shot inference, which is a possible alternative solution compared to FT, because ICL can efficiently adapt to new tasks by learning from given demonstrations. Also, medical institutions can maintain them locally and share limited de-identified data only when needed without sharing all sensitive data for FT. However, the current consensus is that there is a significant performance gap between ICL and FT. Moreover, under the multi-task scenario, FT usually suffers from unbalanced issues, whereas ICL under this setting is underexplored. In this paper, we conduct a comparison between ICL and FT under multi-task setting, exploring their performance gap. Empirical studies show that the advanced ICL method already achieves comparable performance as FT under the multi-task scenario, showing its great potential in medical domains.

1 Introduction

Benefiting from vast learnable model parameters and large-scale pre-training data, large language models (LLMs) achieve a dominant performance in various fields. After conducting multi-task fine-tuning with several high-quality data, LLMs can be further improved (Singhal et al., 2023; Qiu et al., 2024; Singhal et al., 2025) and can solve different downstream tasks with a single model. Such fine-tuning can be considered as instruction-finetuning (Zhang et al., 2023). In medical domains, tons of high-quality real-world data with physicians' annotations are produced from hospitals and medical research institutions every day. Figure 1 shows an example extracted from the MRNER-Disease dataset[1]. However, considering the medical ethical issues such as the privacy of patients and clinical trial participants, it is difficult to borrow these data for fine-tuning LLMs. Meanwhile, it is also risky to do so because LLMs can memorize detailed information in real-world data, especially when LLMs are large enough (Huang et al., 2022; Kiyomaru et al., 2024; Satvaty et al., 2024). Therefore, recent progress on medical LLMs usually rely on open-accessed medical academic papers (Labrak et al., 2024; Wu et al., 2024), clinical guidelines (Chen et al., 2023), medical textbooks (Wang et al., 2024), etc., unabling accessing real-world data from hospitals and medical research institutions.

 # MRNER-Disease
Radiology Report:
右肺下葉S6に境界不明瞭な約12mmのすりガラス状結節を認めます。1年前の画像と比較して、約9mm→約12mmと明らかに増大しています。微小浸潤部分は明瞭でありませんが、微小浸潤性腺癌（MIA）を疑います。
(A vaguely defined ground-glass nodule of approximately 12mm is observed in the S6 section of the bottom right lung lobe. Compared to the image from a year ago, it has clearly enlarged from approximately 9mm to 12mm. The microinvasive part is not clear, but microinvasive adenocarcinoma (MIA) is suspected.)
Abnormal Findings: すりガラス状結節 *(ground-glass nodule)*

Figure 1: A medical named entity recognition example sampled from MRNER-Disease dataset, including a radiology report and human-annotated abnormal findings.

Instead of memorizing the downstream data of training sets inside model parameters, in-context learning (ICL) has been proposed as an alternative way to utilize rich information from a limited number of training samples (Dong et al., 2024). Recent studies show that ICL can be considered as an implicit gradient update on model parameters (Dai et al., 2023; Deutch et al., 2024). Mosbach et al. (2023) show that when the number of training samples is limited, few-shot ICL performs similarly to few-shot FT and they have similar generalization

[1] https://github.com/sociocom/JMED-LLM

ability in out-of-domain downstream tasks. Bertsch et al. (2024) use long-context LLMs (e.g., an LLM fine-tuned with 80k context) to explore the case when the entire training set can be fit into the input context. They find that long-context ICL using the entire training set as demonstrations often approaches or exceeds parameter-efficient fine-tuning (PEFT) on the same scale dataset. However, long-context ICL requires several times more computation in the inference stage than fine-tuned models (using more than 1k demonstrations), which is infeasible in practice, and needs to access the entire training set as FT.

Due to the big success of ICL in adapting LLMs to new tasks, many researchers are dedicated to the development of ICL methods (Dong et al., 2024), including demonstration selection (Rubin et al., 2022; Li and Qiu, 2023), ordering (Liu et al., 2024), etc. Especially, KATE (Liu et al., 2022) retrieves relevant samples from training sets to serve as demonstrations, improving the performance compared to trivial few-shot learning (i.e., random sampling). Existing works (Mosbach et al., 2023; Bertsch et al., 2024) compare ICL and FT under controlled same sample size (few-shot or full-size), but the most realistic setting, namely, comparing few-shot ICL and full-size FT, especially under the multi-task scenario, has been overlooked. In real situations in hospitals, different patients have different presentations of diseases with different requirements for diagnosis, where unbalanced issues are very common. By utilizing the latest few-shot sample selection techniques, we bridge this gap to answer the following research question: **whether current ICL techniques can achieve better or comparable performance compared to standard FT under the multi-task setting**.

In this work, we conduct a comparison between fine-tuning and in-context learning under the multi-task scenario. Our contributions are three-fold:

- To our best knowledge, it is the first work comparing fine-tuning and few-shot in-context learning under the multi-task scenario.
- Empirical studies show that using advanced ICL techniques like KATE, LLMs can recover much improvement from zero-shot to FT.
- This work sheds light on a different path to develop LLMs for medical domains, reducing the risk of exposing sensitive clinical data.

2 Methodology

To study whether current ICL techniques can achieve better or perform comparably to FT, we conduct a relatively fair comparison between them. In the common development process, we usually collect data from downstream tasks and fine-tune the foundation LLMs. In this paper, we perform fine-tuning similarly using the entire available training set. As for ICL, we follow the simple but efficient data selection method KATE (Liu et al., 2022) to perform the comparison. Given a query, KATE aims to search the most similar demonstrations from the candidate set, usually, the training set, as the few-shot demonstrations. Considering computational efficiency during fine-tuning, many parameter-efficient fine-tuning (PEFT) methods have been proposed (Han et al., 2024). By sacrificing some downstream performances, PEFT methods allow us to fine-tune LLMs with less GPU memory. Though their success in saving computation resources, we only compare ICL with full-parameter fine-tuning here, exploring how well the current ICL techniques can be.

In the comparison of this work, we focus on a multi-task scenario, since it is closer to the real-world application involving multiple tasks. We perform fine-tuning on LLMs with training sets containing downstream data from multiple tasks, while we perform KATE retrieving demonstrations from the same corpus. After fine-tuning, we evaluate the fine-tuned models under the zero-shot setting, since the models have already read all training data for multiple epochs. Detailed implementations can be found in Appendix A.1.

3 Experimental Setup

In medical domains, JMedBench is an extensive benchmark including five tasks and 20 datasets in Japanese (Jiang et al., 2025), including multi-choice question-answering (MCQA), named entity recognition (NER), machine translation (MT), document classification (DC), and semantic text similarity (STS). Therefore, it is an ideal testbed for comparing FT and ICL under multi-task scenario. Details of each dataset in JMedBench can be found in Appendix B. We mix all training sets in JMedBench as the fine-tuning corpus and demonstration pool for ICL, resulting in 250,343 training samples. Note that this mixed corpus is unbalanced since the MCQA task has 204k samples, whereas the MT task only contains 80 training samples. As for

experimental subjects, we selected four models including Llama-2-7B (Touvron et al., 2023), Llama-3-8B (Dubey et al., 2024), Qwen-2-7B (Yang et al., 2024), and llm-jp-v3-13B[2]. Details of these LLMs can be found in Appendix A.2.

4 Results and Discussion

4.1 Main Results

Models	MCQA	CMCQA	NER	MT	DC	STS
	Acc (%)		F1 (%)	BLUE	Acc (%)	Pearson
Llama-2-7B	25.83	63.10	27.19	9.52	39.27	-0.2714
+ Standard (3-shot)	27.24	57.90	26.89	**17.78**	**40.46**	0.1121
+ KATE (3-shot)	32.23	57.80	27.19	12.83	39.31	0.2457
+ FT	**37.92**	**71.70**	**39.15**	7.05	37.39	**0.8004**
Llama-3-8B	31.12	55.40	31.50	17.95	39.26	-0.128
+ Standard (3-shot)	35.86	63.10	43.08	24.04	49.48	0.3872
+ KATE (3-shot)	41.20	64.20	40.80	**24.06**	52.56	0.4451
+ KATE (8-shot)	42.28	64.70	**45.69**	21.36	**54.90**	0.5381
+ FT	**43.20**	**71.90**	39.48	5.58	47.89	**0.7926**
Qwen-2-7B	39.69	55.20	23.51	14.73	44.14	-0.007
+ Standard (3-shot)	46.95	59.50	45.07	21.26	56.89	0.5740
+ KATE (3-shot)	49.64	57.80	42.62	**23.72**	55.98	0.5851
+ KATE (8-shot)	49.93	59.70	**48.59**	18.25	**57.35**	0.6441
+ FT	**52.45**	**74.70**	26.51	4.30	53.48	**0.8464**
llm-jp-v3-13B	31.18	73.00	26.51	19.03	36.99	-0.1273
+ Standard (3-shot)	34.80	73.60	39.14	**28.33**	46.25	0.1543
+ KATE (3-shot)	40.17	73.90	36.80	23.91	**47.34**	0.3020
+ KATE (8-shot)	40.60	**74.40**	**40.12**	25.37	36.87	0.3741
+ FT	**45.56**	73.70	37.93	7.05	38.81	**0.8510**

Table 1: Benchmark results on JMedBench. The best and second-best performances when using the same foundation model are highlighted in bold and underlined, respectively.

Sometimes few-shot ICL can outperform full-size FT. As shown in Table 1, although multi-task FT performs better than few-shot ICL in MCQA and STS tasks, when using Llama-3-8B and Qwen-2-7B as the foundation models, using ICL techniques can significantly outperform fine-tuned models in NER, MT, and DC tasks. We believe there are three main reasons. Firstly, JMedBench only contains 80 training samples for the MT task. It is enough for few-shot ICL, however, fine-tuned models may be underfitting. Secondly, unbalanced issues are very common in the multi-task scenario (80 MT samples versus 203k MCQA samples), even resulting in a degradation on the MT task. Thirdly, each NER task has a different annotation schema like granularity. Fine-tuning in a multi-task way may cause a conflict, confusing the models.

How much KATE can recover improvement from zero-shot to multi-task FT? We define the recovery rate α as follows:

$$\alpha_{KATE \to FT} = \frac{P_{KATE} - P_{vanilla}}{P_{FT} - P_{vanilla}} \quad (1)$$

[2] https://huggingface.co/llm-jp/llm-jp-3-13b

Though fine-tuned models outperform models with KATE in the MCQA task, which is an important task in medical domains, ICL may have been underestimated recently. After fine-tuning, Llama-3-8B achieves the best performance in the MCQA task. It is worth noting that with eight retrieved demonstrations, Llama-3-8B using KATE technique recovers 92.38% of improvement from the vanilla model to the multi-task fine-tuned model. Here, we have an inspiring finding that even if we only access the clinical data by retrieving eight relevant demonstrations from the database maintained locally by hospitals, we can still obtain a similar performance on the MCQA task, just like we bring all (maybe anonymized) clinical data out of hospitals for fine-tuning. As for other models like Qwen-2-7B and llm-jp-v3-13B, KATE with 8-shot can also recover 80.25% and 65.51% improvement, respectively.

Better ICL ability, more improvement can be recovered. Llama-3 adopts an attention mask to prevent unexpected attention between different documents within the same sequence (Dubey et al., 2024), which can improve the ICL ability of LLMs (Zhao et al., 2024), learning from the context better. Therefore, Llama-3 has a better ICL ability than Llama-2, which is also consistent with the conclusions of Chen et al. (2025). From Table 1, we notice that Llama-2-7B with KATE only recovers 52.94% improvement under 3-shot evaluation, whereas Llama-3-8B recovers 83.44% improvement. This observation shows that LLMs with KATE technique can perform closer to fine-tuned models when they have better ICL abilities, which illustrates the importance of improving the ICL ability when developing LLMs in the future.

4.2 In-depth Analysis

Accuracy (%)	IGA	JMM	MedM	USM	MedQ	MML
Llama-3-8B	26.31	34.46	32.20	30.87	25.22	37.63
+ Standard (3-shot)	35.31	36.59	35.88	34.01	29.07	44.31
+ KATE (3-shot)	**36.19**	42.33	**49.92**	38.26	32.13	48.37
+ FT	34.94	**42.64**	48.82	**44.93**	**39.12**	**48.74**
Qwen-2-7B	41.81	44.53	35.76	37.71	29.77	48.53
+ Standard (3-shot)	50.94	50.83	42.79	42.18	35.35	**59.59**
+ KATE (3-shot)	**52.06**	51.93	51.14	44.62	38.65	59.43
+ FT	46.19	**52.01**	**55.27**	**53.65**	**48.31**	59.27
llm-jp-v3-13B	28.44	37.84	30.91	29.54	25.22	35.11
+ Standard (3-shot)	37.00	38.87	32.68	33.31	27.02	39.93
+ KATE (3-shot)	39.13	42.01	45.16	39.20	32.76	42.76
+ FT	**45.81**	**44.85**	**48.86**	**43.21**	**40.38**	**50.24**

Table 2: Benchmark results on Japanese biomedical MCQA tasks, including IgakuQA (**IGA**) and JMMLU-medical (**JMM**), MedMCQA-JP (**MedM**), USMLE-QA-JP (**USM**), MedQA-JP (**MedQ**), MMLU-medical-JP (**MML**), and PubMedQA-JP (**Pub**). The best performances are highlighted in bold.

Table 2 shows the performance on six MCQA datasets. We would like to understand where the improvement of KATE comes from and reveal where the gap lies between KATE and FT.

Multi-task KATE fills a gap from standard ICL to multi-task FT. Despite the success of standard ICL, namely, randomly sampling demonstrations for ICL, it mainly understands the task format instead of learning input-output mapping, as suggested by Min et al. (2022). Additionally, KATE adopts retrieved relevant demonstrations, allowing LLMs to learn extra knowledge within the context. As shown in Table 2, with 3-shot demonstrations, LLMs achieve general improvement on all MCQA datasets. Especially, with retrieved demonstrations, KATE achieves further improvement on the MedMCQA-JP dataset no matter which LLM we evaluate. As suggested by Pal et al. (2022), MedMCQA mainly measures acquired knowledge from the LM itself. Therefore, retrieved demonstrations serve as extra knowledge for LLMs. With fine-tuning, LLMs memorize the task format and medical knowledge so that they achieve improved performance under zero-shot evaluation.

Multi-task KATE overcomes the obstacle when in-domain training samples are insufficient. Besides MedMCQA-JP, multi-task KATE achieves larger improvement on JMMLU-Medical and MMLU-Medical-JP datasets, which contain only 45 training samples. Standard ICL only selects demonstrations in the in-domain training set. With a retriever, multi-task KATE allows selecting demonstrations across different datasets. Although the retriever may select demonstration from other tasks (e.g., retrieved NER sample when completing MCQA task), Table 2 shows that multi-task KATE benefits more from the larger candidate set.

Multi-task KATE is limited when reasoning is required. USMLE-QA-JP and MedQA-JP are two datasets derived from medical license examinations, containing complex questions describing real clinical scenarios. They require clinical reasoning ability, such as analyzing differential diagnoses and finding optimal treatment options. On these two datasets, despite multi-task KATE outperforms standard ICL, gaps between KATE and FT are difficult to close. We believe it is because retrieved demonstrations provide information on task format and related knowledge, however, LLMs cannot do proper reasoning based on them. How to improve the reasoning capability of LLMs during ICL is still challenging nowadays.

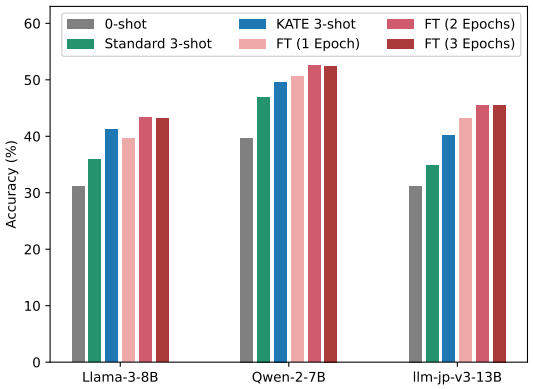

Figure 2: Performances on the MCQA task of different settings using different foundation LLMs.

LLMs learn complex patterns after 1 epoch of fine-tuning. Figure 2 shows the performances on the MCQA task of different settings using different foundation LLMs. We find that KATE performs closer to LLMs fine-tuned by 1 epoch, for example, Qwen-2-7B with three retrieved demonstrations recovers 91.71% of FT improvement and similarly llm-jp-v3-13B recovers 74.79% of FT improvement. Furthermore, KATE can even outperform LLMs fine-tuned with 1 epoch, for example, Llama-3-8B with three retrieved demonstrations achieves 1.61% absolute accuracy improvement over Llama-3-8B fine-tuned with 1 epoch.

5 Conclusions

In this paper, we perform a comparison between the advanced in-context learning technique, KATE, and full-parameter fine-tuning under the multi-task setting using the entire training sets. Experimental results on the JMedBench show that current ICL techniques may be underestimated in medical domains. With several retrieved demonstrations, KATE allows LLMs to recover a large proportion of improvement from vanilla LLMs to fine-tuned LLMs, for example, Llama-3-8B can recover 92.38% improvement. Therefore, in the future, if we could develop better ICL methods or LLMs with better ICL ability, obtaining the entire training dataset from medical institutions for fine-tuning is unnecessary. Instead, the medical institutions can de-identify and maintain their data locally, while we access a small proportion of this database by retrieving relevant clinical data, leading to a similar performance. We hope this work can motivate future research on developing better ICL techniques to achieve comparable or even better performance than FT as well as improving the ICL ability of foundation LLMs in medical domains.

Limitations

In this work, the ICL technique for competing with FT has a large room for improvement. Even though we adopt simple deduplication on the ICL demonstrations, retrievers easily retrieve highly similar data, which cannot provide extra information. Therefore, how to remove those data and improve the diversity of selected demonstrations is a promising way to improve the performance of ICL methods in medical domains in the future.

Our methodology is not hinged to the Japanese language. If experiments on other languages like English, Chinese, and French were available, the conclusions of this work would be more solid. However, in medical domains, current LLM researchers focus mainly on the MCQA task. It is difficult to find a proper benchmark with multiple tasks. If benchmarks with multiple tasks in other languages are available, experiments on such benchmarks should be done.

Besides, during our experiments, we realized that the quality of some translated training samples may not be satisfactory enough for fine-tuning, which may limit the performance of fine-tuning, although they may not effect the performance of in-context learning. In the future, to further confirm the conclusions drawn from this work, experiments on higher quality training corpus are required.

Ethics Statement

We follow the statement of the JMebBench and open-sourced LLMs including Llama-2, Llama-3, Qwen-2, and llm-jp-v3 carefully in our experiments.

Our experimental results have limitations because of the limited datasets, tasks, and models. Though we show that advanced ICL methods can achieve comparable or even better performance than FT, it should be noted that LLMs with retrieved demonstrations from hospitals or medical research institutions should be treated carefully. Such models can still generate unfaithful content even though relevant contexts are given. Therefore, those who want to use similar techniques to develop faithful biomedical LLM-based applications should be aware of this limitation.

References

Amanda Bertsch, Maor Ivgi, Uri Alon, Jonathan Berant, Matthew R Gormley, and Graham Neubig. 2024. In-context learning with long-context models: An in-depth exploration. *arXiv preprint arXiv:2405.00200*.

Wentong Chen, Yankai Lin, ZhenHao Zhou, HongYun Huang, YanTao Jia, Zhao Cao, and Ji-Rong Wen. 2025. ICLEval: Evaluating in-context learning ability of large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10398–10422, Abu Dhabi, UAE. Association for Computational Linguistics.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

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

Gilad Deutch, Nadav Magar, Tomer Natan, and Guy Dar. 2024. In-context learning and gradient descent revisited. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1017–1028, Mexico City, Mexico. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

P Goyal. 2017. Accurate, large minibatch sg d: training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*.

Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. 2022. Are large pre-trained language models leaking your personal information? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2038–2047, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning.

Junfeng Jiang, Jiahao Huang, and Akiko Aizawa. 2025. JMedBench: A benchmark for evaluating Japanese biomedical large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5918–5935, Abu Dhabi, UAE. Association for Computational Linguistics.

Hirokazu Kiyomaru, Issa Sugiura, Daisuke Kawahara, and Sadao Kurohashi. 2024. A comprehensive analysis of memorization in large language models. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 584–596, Tokyo, Japan. Association for Computational Linguistics.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. BioMistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5848–5864, Bangkok, Thailand. Association for Computational Linguistics.

Xiaonan Li and Xipeng Qiu. 2023. Finding supporting examples for in-context learning. *CoRR*.

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.

Yinpeng Liu, Jiawei Liu, Xiang Shi, Qikai Cheng, Yong Huang, and Wei Lu. 2024. Let's learn step by step: Enhancing in-context learning ability with curriculum learning. *arXiv preprint arXiv:2402.10738*.

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.

Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar. 2023. Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12284–12314, Toronto, Canada. Association for Computational Linguistics.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Towards building multilingual language model for medicine. *Nature Communications*, 15(1):8384.

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.

Ali Satvaty, Suzan Verberne, and Fatih Turkmen. 2024. Undesirable memorization in large language models: A survey. *arXiv preprint arXiv:2410.02650*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8.

Larry Smith, Lorraine K Tanabe, Rie Johnson nee Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M Friedrich, Kuzman Ganchev, et al. 2008. Overview of biocreative ii gene mention recognition. *Genome biology*, 9:1–19.

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*.

Xidong Wang, Nuo Chen, Junyin Chen, Yan Hu, Yidong Wang, Xiangbo Wu, Anningzhe Gao, Xiang Wan, Haizhou Li, and Benyou Wang. 2024. Apollo: Lightweight multilingual medical llms towards democratizing medical ai to 6b people. *arXiv preprint arXiv:2403.03640*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen,

Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, page ocae045.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Yu Zhao, Yuanbin Qu, Konrad Staniszewski, Szymon Tworkowski, Wei Liu, Piotr Miłoś, Yuxiang Wu, and Pasquale Minervini. 2024. Analysing the impact of sequence composition on language model pretraining. *arXiv preprint arXiv:2402.13991*.

A Experimental Details

A.1 Implementation details

All the experiments were implemented mainly based on PyTorch (Paszke et al., 2019) and transformers (Wolf et al., 2020). Model checkpoints were downloaded from Huggingface[3] and the corresponding checkpoints are listed in Table 3.

When performing fine-tuning, we refer to the hyperparameter settings from the original papers when the authors developed their instructed version of LLMs. Since we only use two nodes with 8 NVIDIA A100 GPUs for fine-tuning llm-jp-v3-13B and a single node for fine-tuning the rest of the involved comparison methods, we follow Goyal (2017) to adjust the hyperparameters correspondingly. Some important hyperparameters are summarized in Table 4. During fine-tuning, we adopt a warmup strategy in 10% of total steps at the beginning and decrease the learning rate gradually to 10% of the peak learning rate.

[3] https://huggingface.co/models

Model	Checkpoint
Llama-2-7B	meta-llama/Llama2-7b-hf
Llama-3-8B	meta-llama/Meta-Llama3-8B
Qwen-2-7B	Qwen/Qwen2-7B
llm-jp-v3-13B	llm-jp/llm-jp-3-13b

Table 3: The corresponding checkpoints in the model hub of Huggingface for involved comparison methods.

Model	LR	Global BS
Llama-2-7B	2e-5	64
Llama-3-8B	2e-5	64
Qwen-2-7B	1.4e-5	64
llm-jp-v3-13B	1.25e-5	64

Table 4: Hyperparameters for fine-tuning. LR: Learning Rate. BS: Batch size.

When performing standard ICL, we tried to do random sampling from the entire mixed training set of JMedBench in our preliminary experiment. However, since the sampled demonstrations were probably not from the same task, they could not help LLMs to predict, performing close to zero-shot performance as vanilla LLMs. Therefore, in this work, we randomly sample demonstrations from the corresponding in-domain training set instead of the whole corpus.

When applying the KATE technique, we chose multilingual Contriever (Izacard et al., 2021) as our retriever. We retrieved three or eight similar training samples from the mixed training sets of JMedBench as demonstrations for ICL. Note that it is not guaranteed that every testing sample retrieves demonstrations from the same task. For example, when evaluating the IgakuQA dataset, which belongs to MCQA tasks, some NER samples will be included. We tried to remove them from the demonstration candidate set, however, there was no significant difference. We hypothesize that data retrieved from the different tasks can also help to provide extra knowledge for prediction, just as multi-task fine-tuning does.

A.2 Details of Experimental Subjects

Llama-2-7B and Llama-3-8B are two versions of the Llama model, which were pre-trained mainly

in English. Considering significant overlapping tokens in Chinese and Japanese, we include Qwen-2-7B as an experimental subject, which is a multilingual LLM pre-trained mainly in Chinese and English. llm-jp-v3-13B is a representative Japanese LLM, which is suitable for analyzing on Japanese tasks.

A.3 Prompt Engineering

For the sake of simplification, we only use the Standard template for each task as suggested by Jiang et al. (2025).

B Details of the JMedBench

JMedBench contains five medical tasks, including multi-choice question-answering (MCQA), named entity recognition (NER), machine translation (MT), document classification (DC), and semantic text similarity (STS). Besides human-handcrafted Japanese medical data, the authors translated some large-scale, high-quality medical datasets in English, such as MedMCQA (Pal et al., 2022) and BC2GM (Smith et al., 2008). Table 5 shows the statistics of this benchmark. Further details can be found in the original paper (Jiang et al., 2025).

Task	Dataset	Train	Test
MCQA	IgakuQA	10,178	989
	JMMLU-medical	45	1,271
	MedMCQA-JP	182,822	4,183
	USMLE-QA-JP	10,178	1,273
	MedQA-JP	10,178	1,273
	MMLU-medical-JP	45	1,871
	PubMedQA-JP	1,000	1,000
MT	EJMMT	80	2,400
NER	MRNER-Medicine	10	90
	MRNER-Disease	10	90
	NRNER	10	90
	BC2GM-JP	12,572	5,037
	BC5Chem-JP	4,562	4,801
	BC5Disease-JP	4,560	4,797
	JNLPBA-JP	18,607	4,260
	NCBI-Disease-JP	5,424	940
DC	CRADE	8	92
	RRTNM	11	89
	SMDIS	16	84
STS	JCSTS	170	3,500

Table 5: Statistics of datasets in JMedBench. PubMedQA-JP includes an extra abstract. We analyze it separately in our main experiments and abbreviate it as CMCQA (Context-based MCQA).