Dr.ICL: Demonstration-Retrieved In-context Learning

Anonymous Author(s)

Abstract

In-context learning (ICL), which teaches a large language model (LLM) to perform a task with few-shot demonstrations rather than adjusting the model parameters, has emerged as a strong paradigm for using LLMs. While early studies primarily used a fixed or random set of demonstrations for all test queries, recent research suggests that retrieving semantically similar demonstrations to the input from a pool of available demonstrations results in better performance. This work expands the applicability of retrieval-based ICL approaches along several dimensions. We extend the success of retrieval-based ICL to instruction-finetuned LLMs as well as Chain-of-Thought (CoT) prompting. While the prior work utilizes general Large Language Models (LLMs), such as GPT-3, we find that retrieved demonstrations also enhance instruction-finetuned LLMs. This insight implies that training data, despite being exposed during the fine-tuning phase, can still be effectively used through retrieval and in-context demonstrations during testing, resulting in superior outcomes when compared to utilizing no demonstrations or selecting them at random. For CoT, when the demonstrations contain reasoning chains, we get improvements by retrieving based on such chains. Finally, we train a task-specific demonstration retriever that outperforms off-the-shelf retrievers.

1 Introduction

Language models are now the foundation models for many natural language processing tasks across a wide range of domains [2]. One of the most exciting emergent abilities [27] of large language models (LLMs) is in-context learning (ICL) [3]. With ICL, instructions and a few demonstrative examples are augmented to the inputs of LLMs, allowing them to perform well on new tasks without the need for fine-tuning.

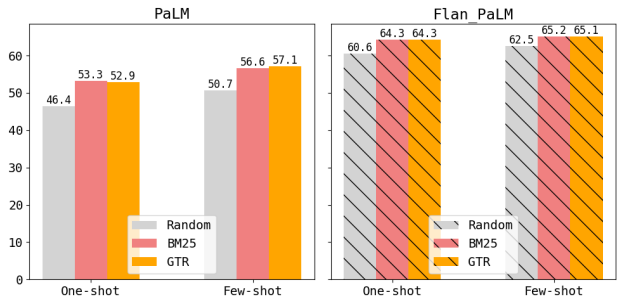

Figure 1: The average performance of PaLM and Flan-PaLM on five datasets, with one and few-shot ICL. Retrieved demonstrations given by either BM25 or GTR yield better performance than random demonstrations.

Typically, ICL approaches utilize random or hand-crafted demonstrations that are applied across various queries. This may, however, not always be optimal. Recent research has revealed that using

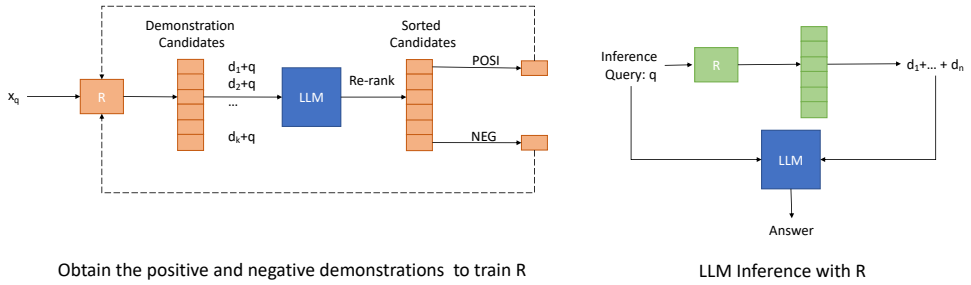

Figure 2: Pipeline for training demonstration retriever and inference (R for a neural retriever). Figure on the left shows the procedure of obtaining data to train a demonstration retriever: an off-the-shelf retriever takes an input query x_q and retrieves top-k (e.g., 100) demonstrations candidates from the training corpus. Then an LLM is used to output the score of the ground truth of y_q with each retrieved demonstration and x_q. Figure on the right shows the inference pipeline for in-context learning with the trained demonstration retriever.

demonstrations semantically similar to the input query can enhance performance [14]. Here, we investigate two off-the-shelf retrievers, BM25 [23] and GTR [20], where BM25 is a sparse retriever that finds demonstrations with the highest (weighted) word overlap with the query, while GTR is a dense retriever that seeks demonstrations semantically closest to the query. We use these retrievers to obtain query-specific demonstrations, and study demonstration-retrieved ICL (Dr. ICL) with both a general LLM and an instruction-finetuned LLM.

Beyond previous work, several interesting findings are discovered through our experiments, as shown in Figure 1. Firstly, despite their simplicity, we establish that both BM25 and GTR can find more effective demonstrations than random demonstrations in both one-shot and few-shot ICL settings. Such off-the-shelf retrievers make Dr. ICL an appealing paradigm for real-world applications. Secondly, our results with an instruction-finetuned LLM, i.e., Flan-PaLM [7], indicate that training data can be useful not only for training models but for accompanying a retriever at inference time, suggesting a more efficient way to utilizing training data which are expensive to collect. Thirdly, the combination of Chain-of-Thought (CoT) [26] and retrieved demonstrations surpasses the performance of CoT alone. Moreover, selecting demonstrations based on their annotated reasoning chains proves to be more beneficial than retrieving without considering reasoning chains. Last but not least, we provide a detailed analysis by comparing retrieved demonstrations with input queries. This analysis underscores an interesting observation: the labels of the retrieved demonstrations frequently do not coincide with the input query's label across multiple tasks. We also compare the diversity of demonstrations and find that diversity might impact how much improvement Dr.ICL makes.

Nevertheless, while demonstrations from off-the-self retrievers perform better than random ones, they may still be suboptimal for the target task, since the retrievers were optimized for other tasks such as question answering and information retrieval. We thus propose to train a demonstration retriever that retrieves the most beneficial demonstrations for the target task as demonstrated in Figure 2 (§2.2). Our experimental results show that the fine-tuned demonstration retriever outperforms off-the-shelf retrievers, with more noticeable improvement in one-shot ICL. This encouraging result indicates that for a fixed target task, the trained retriever could offer an effective substitute for off-the-shelf models. Given the constraints of the page limit, we've provided the related work and a comparison with prior studies in AppendixA.

2 Demonstration-Retrieved In-Context Learning (Dr. ICL)

We first describe ICL for general tasks (including both classification and generation tasks). For a task T, given an input text x_q, an LLM is used to predict the answer y_q conditioned on a set of *demonstrations*, $Demo = \{d_1, d_2, \cdots, d_n\}$, where $d_i = (x_i, y_i)$ is a pair of input and ground truth answer. Typically, d_i is linearized as a string (e.g., "`question:` x_i `\n answer:` y_i") and then provided to the LLM alongside x_q.

There are multiple strategies for choosing the set of demonstrations. For instance, one could randomly or manually select a fixed set $Demo$ to be applied to all queries of task T. Alternatively, a retriever can be used to find query-specific demonstrations from the training set D_{train}:

$$Demo_{x_q} = Retriever(x_q, D_{train}, n), \tag{1}$$

where $Demo_{x_q}$ contains the top-n demonstrations that the retriever considers most suitable for the input x_q. In this work, we consider two off-the-shelf retrievers, BM25 and GTR (Section 2.1), and then propose a method to train a retriever tailored to the target task T (Section 2.2).

2.1 Off-the-shelf Retrievers

BM25 [23] is a bag-of-words model that calculates relevance scores using term frequency, inverse document frequency, and document length normalization. It has proven effective and efficient, making it easily deployable in large-scale, real-world applications. However, BM25 heavily relies on keyword matching and lacks context understanding, which may result in less accurate outcomes. In contrast, GTR [20] is a dual-encoder neural retriever (based on T5) trained on the MS MARCO dataset [19]. GTR excels in semantic and context comprehension and is easily transferable to downstream tasks or specific domains. However, it has lower memory and computational efficiency, and lacks interpretability.

2.2 Demonstration Retriever Training

Demonstration retrieval aims to find the most representative demonstrations for each input query. Ideally, the demonstrations should capture both (a) the query-specific knowledge required to answer the query, and (b) the nature of the task and how the task should be solved in general.

Off-the-shelf retrievers such as BM25 and GTR were designed for information retrieval and question answering. As such, they mostly retrieve demonstrations of type (a) but not (b). To fill this gap, we propose to train a demonstration retriever by leveraging the feedback from a language model. As demonstrated in the left part of Figure 2, the process involves two steps: obtaining the training data and training a retriever on the data.

Obtain the Training data We want to teach the retriever model to locate examples that lead to the most accurate predictions. We propose to mine a set of demonstrations for each input query x_q in the training data as follows. First, given a question-answer pair $(x_q, y_q) \in D_{train}$, we use an off-the-shelf retriever to find a demonstration candidate set D for x_q, where x_q itself is excluded from D. We then test each demonstration $d \in D$ on how much it helps on the target task. Specifically, the LM probability $p_{\text{LM}}(y_q \mid d, x_q)$ of the gold answer y_q is used as the score for the demonstration. Finally, we keep the top-n demonstration as the positive demonstrations, and the bottom-n as the hard negative demonstrations.

Training Procedure Our retriever is a dual encoder, which defines the score of any query-document pair (q, d) as $s(q, d) = v_q^\top v_d$, where v_q and v_d are the embeddings of q and d. We initialize our retriever with GTR, and then fine-tune it on the training data via contrastive loss with both in-batch and hard negatives:

$$\mathcal{L}_{con} = -\log \frac{e^{s(q,d^+)}}{e^{s(q,d^+)} + \sum_j e^{s(q,d_j^-)}}, \tag{2}$$

where d^+ and d_j^- are the positive and negative demonstrations. The negative demonstrations include the positive demonstrations for the other input queries in the same batch and one randomly-chosen hard negative demonstration.

3 Experiments

Benchmarks We study various tasks across 5 datasets: free-form question answering (NQ), natural language inference (ANLI-r3), mathematical reasoning (GSM8K and AQuA) and boolean question answering (StrategyQA). All the tasks are evaluated by exact match accuracy.

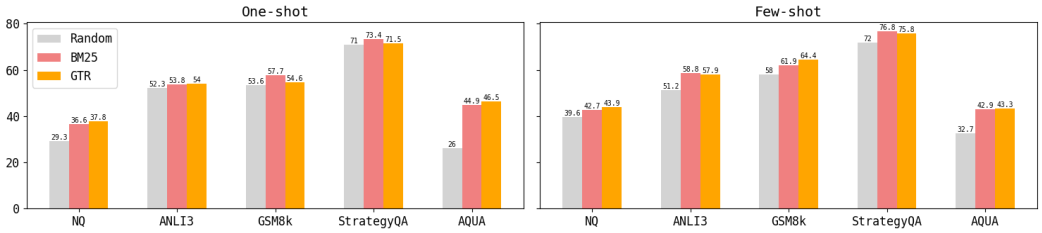

Figure 3: PaLM: One-shot and few-shot inference with three types of demonstrations: random, BM25, and GTR. Retrieved demonstrations are more effective than random ones.

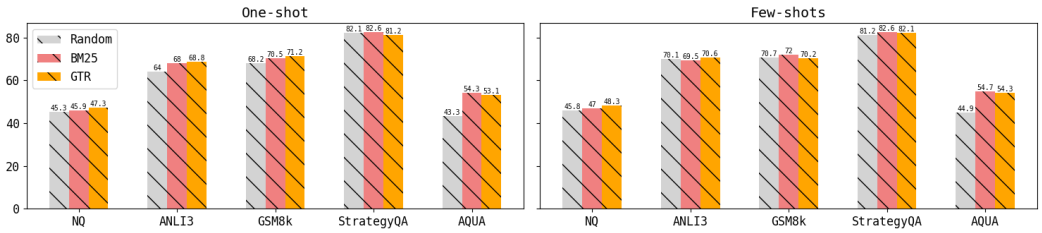

Figure 4: Flan-PaLM: One-shot and few-shot inference with three types of demonstrations: random, BM25, and GTR. Retrieved demonstrations are more effective than random ones.

Language Models PaLM-540B [6] and Flan-PaLM (540B) [7] are used as the primary LLMs. Both models have the same architecture, but Flan-PaLM has been further trained on thousands of tasks for instruction learning (including all the five datasets we studied in this paper) and shows superior generalization performance compared to PaLM. For GSM8K, AQuA, and StrategyQA, we also apply Chain-of-Thought (CoT) prompting [28], which has shown effectiveness in such complex reasoning tasks. The main idea is to have the LLM generate a CoT containing reasoning steps before generating the answer. In order to induce such a behavior, each in-context demonstration is additionally equipped with a CoT, which is available from the training data. Note that the CoT can also be utilized during retrieval, and in our experiments, we will show the benefits of retrieving based on the CoT. We use the temperature of 0.0 and maximum decoding length 10 for tasks without CoT and 256 for tasks involving CoT.

Retrievers As explained in §2, we explore using BM25 and GTR as off-the-shelf retrievers, as well as training our own retriever for each task. For BM25, we use uncased BERT wordpiece tokenization and parameters $(k_1, b) = (1.5, 0.75)$. For GTR, we use the pretrained GTR-Base model. When mining data for training our retriever, we use the pretrained GTR to retrieve 100 demonstrations candidates, and then use PaLM-62B to score each candidate. (We used the smaller PaLM-62B instead of 540B for efficiency.) Then we select the top-5 reranked demonstrations as the positive candidates to fine-tune GTR. **Retrieval Corpus.** We create a separate retrieval corpus for each task using the associated training data. For tasks with CoT, each entry in the corpus is composed of the question, the CoT, and the answer, while for other tasks are without the CoT.

3.1 Off-the-shelf-retriever Performance

Figures 3 and 4 show the performance of PaLM and Flan-PaLM under one-shot and few-shot ICL settings, with and without retrievers. We make the following observations.

Observation 1: Off-the-shelf retrievers are capable of finding more effective demonstrations than random ones. Figure 3 shows that the demonstrations retrieved by BM25 or GTR are better than random ones under both one-shot and few-shot scenarios for the PaLM model, suggesting that in real life applications, rather than using random demonstrations, retrieved demonstrations can yield better performance. It is worth mentioning that BM25 is more efficient in terms of indexing memory and retrieval latency compared to semantic retrievers like GTR or other sentence encoders [14], which makes it easier to deploy.

# Shots	w/o CoT	PaLM			Flan-PaLM		
		GSM8K	StrategyQA	AQuA	GSM8K	StrategyQA	AQuA
One	without	47.2	62.8	42.5	65.4	78.3	51.2
	with	**57.7**	**73.4**	**44.9**	**70.5**	**82.6**	**54.3**
Few	without	58.5	72.9	**44.1**	71.3	80.2	54.0
	with	**61.9**	**76.8**	42.9	**72.0**	**82.6**	**54.7**

Table 1: Comparison between two strategies of retrieving demonstrations: with or without CoT. BM25 is used as the retriever. Retrieval with CoT is better than without.

# Shots	w/o CoT	PaLM			Flan-PaLM		
		GSM8K	StrategyQA	AQuA	GSM8K	StrategyQA	AQuA
One	without	49.1	61.8	43.7	66.1	72.5	51.2
	with	**54.6**	**71.5**	**46.5**	**70.5**	**82.6**	**54.3**
Few	without	59.4	70.5	**46.1**	69.9	78.7	54.3
	with	**64.4**	**75.8**	43.3	**70.2**	**82.1**	**54.3**

Table 2: Comparison between two strategies of retrieving demonstrations: with or without CoT. GTR is used as the retriever. Retrieval with CoT is better than without.

Observation 2: Dr. ICL improves instruction-finetuned LLM. Previous research has primarily focused on investigating demonstration retrieved ICL with general LLMs (such as GPT-3) rather than instruction-finetuned LLMs, possibly because they did not consider reusing the training data. In our study, we examine Dr. ICL with Flan-PaLM, an instruction-finetuned LLM, and present the results in Figure 4. Overall, the retrieved demonstrations outperform no demonstrations or random demonstrations. This implies that the training data should be reused during inference as they can be retrieved and enhance the performance, even if the model has already seen such data. We conjecture that the retrieved demonstrations may enhance knowledge localization for ICL, which could explain the observed improvement.

Observation 3: Dr. ICL can further improve advanced prompting technique, Chain-of-Thought. In our experiments on GSM8k, StrategyQA, and AQuA, using Dr. ICL in conjunction with CoT results in improved performance under both one-shot and few-shot ICL scenarios. This finding suggests that Dr. ICL has the potential to enhance the performance of powerful prompting techniques.

The observations above hold significant values for real-world applications. Incorporating ICL with a simple BM25 demonstration retriever, which is highly scalable in terms of latency and indexing memory, is proven to improve the performance of the LLM, including when instruction finetuning or Chain-of-Thought were used. Examples of retrieved demonstrations given by the off-the-shelf retrievers are given in the Table 6 in Appendix.

Retrieval Strategies for CoT For the tasks involving CoT, we evaluate two approaches for retrieving demonstrations: incorporating CoT and excluding CoT when computing the retrieval scores (e.g., for GTR, including CoT means that the demonstration's CoT is add to the input of the embedding model). Note that in both approaches, the CoT will then be added to the in-context demonstrations during LLM inference, so that the LLM knows to generate a CoT. Tables 1 and 2 indicate that implementing CoT in the retrieval phase typically provides better results (with only one exception on the AQuA dataset). This holds true for both one-shot and few-shot scenarios with the PaLM and Flan-PaLM models. Thus, we suggest the integration of CoT during retrieval.

3.2 Trained Demonstration Retriever Performance

We experiment our trained demonstration retriever with PaLM. Table 3 displays both one-shot and few-shot performance, and shows that the trained demonstration retriever is better than off-the-shelf GTR in almost all cases, leading to a better overall performance. Notably, the improvements were most significant in one-shot ICL scenarios, which require less inference latency and computing resources than few-shot ICL. These promising results suggest that the trained retriever could provide an effective alternative to off-the-shelf models.

Task	Method	One Shot	Few Shots
NQ	GTR	37.8	43.9
	Demo-GTR(our)	**39.2(+1.4)**	43.9
ANLI (r3)	GTR	54.0	57.9
	Demo-GTR(our)	**54.8(+0.8)**	**59.0(+1.1)**
GSM8k	GTR	54.6	**64.4**
	Demo-GTR(our)	**59.3(+4.7)**	61.5(-2.9)
StrategyQA	GTR	71.5	75.8
	Demo-GTR(our)	**72.0(+0.5)**	**77.3(+1.5)**
AUQA	GTR	**46.5**	43.3
	Demo-GTR(our)	42.5(-4.0)	**44.5(+1.2)**
Avg.	GTR	52.9	57.1
	Demo-GTR(our)	**53.6(+0.7)**	**57.2(+0.1)**

Table 3: Performance of PaLM using GTR and Demo-GTR retrieved demonstrations. Demo-GTR consistently achieves better performance than GTR in one-shot case.

4 Analysis

To rule out the chance that retrieved demonstrations are more advantageous than random ones simply because in the benchmark datasets the former's answers are identical to the correct ones, we assess the overlap percentage between the demonstration responses and the target. In the few-shot scenario, we aggregate the answers from the demonstrations via majority voting. From Table 4, it is evident that for the first forth datasets, the overlap ratio is roughly equal to or less than the uniform distribution, suggesting that the benefits of the retrieved demonstrations are not due to label identification. In the case of the NQ, we notice a considerable overlap between demonstration answers and the ground truth. We then randomly select 100 instances out of the 433 overlapped cases from GTR-retrieved demonstrations (one-shot) and manually examine them. We find that, indeed, for the majority of the 100 instances, the input questions are semantically equal to the demonstration questions.

Task	Random	Retriever	One-shot	Few-shot
ANLI3	33.33	BM25	33.33	31.42
		GTR	34.75	32.25
StrategyQA	50.0	BM25	48.79	47.34
		GTR	47.83	48.31
AQUA	20.0	BM25	22.83	25.98
		GTR	24.02	22.05
GSM8K	0.0	BM25	1.36	1.82
		GTR	0.99	1.14
NQ	0.0	BM25	8.95	8.70
		GTR	11.99	11.08

Table 4: Overlapped Ratio of Demonstrations Answers with Targets: **Random** represents the probability of selecting the correct label if we select randomly from the space of possible labels.

5 Discussion and Conclusion

In our study, we employed off-the-shelf retrievers to boost ICL through query-oriented demonstrations. Our findings show that these retrievers outperform random demonstrations. Using Flan-PaLM, we highlight that training data enhances fine-tuned LLM performance during ICL testing. Combining Dr. ICL with advanced prompting techniques, as seen in our CoT experiments, further bolsters model performance. We also detail a method to train a demonstration retriever that surpasses off-the-shelf retrievers, especially in one-shot scenarios. Exploring demonstrations across tasks without training data presents a promising future research avenue.

Bibliography

[1] S. Arora, A. Narayan, M. F. Chen, L. J. Orr, N. Guha, K. Bhatia, I. Chami, F. Sala, and C. Ré. Ask me anything: A simple strategy for prompting language models. *arXiv preprint arXiv:2210.02441*, 2022.

[2] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[3] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[4] X. Chen, L. Li, N. Zhang, X. Liang, S. Deng, C. Tan, F. Huang, L. Si, and H. Chen. Decoupling knowledge from memorization: Retrieval-augmented prompt learning. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[5] D. Cheng, S. Huang, J. Bi, Y. Zhan, J. Liu, Y. Wang, H. Sun, F. Wei, D. Deng, and Q. Zhang. Uprise: Universal prompt retrieval for improving zero-shot evaluation. *arXiv preprint arXiv:2303.08518*, 2023.

[6] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

[7] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

[8] D. Dai, Y. Sun, L. Dong, Y. Hao, Z. Sui, and F. Wei. Why can gpt learn in-context? language models secretly perform gradient descent as meta optimizers. *arXiv preprint arXiv:2212.10559*, 2022.

[9] B. Dalvi, O. Tafjord, and P. Clark. Towards teachable reasoning systems. *arXiv preprint arXiv:2204.13074*, 2022.

[10] Y. Fu, H. Peng, A. Sabharwal, P. Clark, and T. Khot. Complexity-based prompting for multi-step reasoning. *arXiv preprint arXiv:2210.00720*, 2022.

[11] H. Gonen, S. Iyer, T. Blevins, N. A. Smith, and L. Zettlemoyer. Demystifying prompts in language models via perplexity estimation. *arXiv preprint arXiv:2212.04037*, 2022.

[12] M. Kazemi, S. Mittal, and D. Ramachandran. Understanding finetuning for factual knowledge extraction from language models. *arXiv preprint arXiv:2301.11293*, 2023.

[13] S. Kumar and P. Talukdar. Reordering examples helps during priming-based few-shot learning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4507–4518, Online, Aug. 2021. Association for Computational Linguistics.

[14] J. Liu, D. Shen, Y. Zhang, W. B. Dolan, L. Carin, and W. Chen. 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, 2022.

[15] X. Lyu, S. Min, I. Beltagy, L. Zettlemoyer, and H. Hajishirzi. Z-icl: Zero-shot in-context learning with pseudo-demonstrations. *arXiv preprint arXiv:2212.09865*, 2022.

[16] A. Madaan, N. Tandon, P. Clark, and Y. Yang. Memory-assisted prompt editing to improve gpt-3 after deployment. *arXiv preprint arXiv:2201.06009*, 2022.

[17] A. Madaan and A. Yazdanbakhsh. Text and patterns: For effective chain of thought, it takes two to tango. *arXiv preprint arXiv:2209.07686*, 2022.

[18] S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*, 2022.

[19] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng. Ms marco: A human generated machine reading comprehension dataset. *choice*, 2640:660, 2016.

[20] J. Ni, C. Qu, J. Lu, Z. Dai, G. H. Ábrego, J. Ma, V. Y. Zhao, Y. Luan, K. B. Hall, M.-W. Chang, et al. Large dual encoders are generalizable retrievers. *arXiv preprint arXiv:2112.07899*, 2021.

[21] E. Nie, S. Liang, H. Schmid, and H. Schütze. Cross-lingual retrieval augmented prompt for low-resource languages. *ArXiv*, abs/2212.09651, 2022.

[22] O. Ram, Y. Levine, I. Dalmedigos, D. Muhlgay, A. Shashua, K. Leyton-Brown, and Y. Shoham. In-context retrieval-augmented language models. *arXiv preprint arXiv:2302.00083*, 2023.

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

[24] O. Rubin, J. Herzig, and J. Berant. 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, 2022.

[25] M. Suzgun, N. Scales, N. Schärli, S. Gehrmann, Y. Tay, H. W. Chung, A. Chowdhery, Q. V. Le, E. H. Chi, D. Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.

[26] B. Wang, S. Min, X. Deng, J. Shen, Y. Wu, L. Zettlemoyer, and H. Sun. Towards understanding chain-of-thought prompting: An empirical study of what matters. *arXiv preprint arXiv:2212.10001*, 2022.

[27] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, et al. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022.

[28] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. Chi, Q. Le, and D. Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

[29] S. M. Xie, A. Raghunathan, P. Liang, and T. Ma. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*, 2021.

[30] J. Ye, Z. Wu, J. Feng, T. Yu, and L. Kong. Compositional exemplars for in-context learning. *arXiv preprint arXiv:2302.05698*, 2023.

[31] X. Ying. An overview of overfitting and its solutions. In *Journal of physics: Conference series*, volume 1168, page 022022. IOP Publishing, 2019.

[32] Z. Zhao, E. Wallace, S. Feng, D. Klein, and S. Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR, 2021.

A Related Work

Paper	LLMs	Retrieval Method	Retrieval Corpus	Evaluation Tasks	# of Shots in Prompts	CoT
KATE [14]	GPT-3	RoBERTa+kNN	In-Domain TD	SA, T2T	Few-shots	No
EPR [24]	GPT-J, GPT-Neo, CODEX, GTP-3	SBERT, BM25, FT Retriever	In-Domain TD	SRM	Few-shots	No
CEIL [30]	GPT-Neo, GPT2-XL, CodeX	BM25, BERT, DPR, FT Retriever	In-Domain TD	SA, PD, NLI, CSR, QA, codeG, and SP	Few shots	No
UPRISE [5]	GPT-Neo, BLOOM, OPT, GPT-3	FT Retriever	Cross Tasks TD	RC, QA, NLI, SA, CSR, CR, PD	Few shots	No
Dr.ICL (Ours)	PaLM, Flan-PaLM	BM25, GTR, FT Retriever	In-Domain TD	QA, NLI, MathR, BC	One-shot, Few-shots	Yes

Table 5: Comparison with Related Work. TD: training data, QA: question answering, RC: reading comprehension, NLI: natural language inference, SA: sentiment analysis, CSR: commonsense reasoning, CR: Coreference Resolution, MathR: mathmatical reasoning, PD: paraphrase detection, SP:semantic parsing, CodeG: code generation, SRM: Sentence representation mapping, T2T: Table to Text generation, Question Answering,

A.1 Few-shot In-context Learning

Few-shot in-context learning (ICL) is a technique that allows language models, such as GPT-3 [3] and PaLM [6], to generalize to new tasks based on a small number of examples. ICL offers several advantages over the traditional training approach of language models, which involves pre-training followed by fine-tuning. One key benefit is that fine-tuning may not always be feasible due to restricted access to the LLM or inadequate computational resources [3]. Additionally, ICL avoids the issues commonly associated with fine-tuning, such as overfitting or shocks [31, 12], as it does not modify the model's parameters, allowing it to remain general.

However, the effectiveness of ICL hinges on various factors, such as the order of the demonstrations [13], the distribution of the demonstrations [18], and the complexity and quality of the prompts themselves [32, 1]. Some research has shown that lower perplexity prompts [11] and open-ended question-answer formats [1] tend to lead to better performance, while others have found that intermediate reasoning steps [28] and higher complexity prompts [10] can also improve results on certain tasks [25, 26]. In an effort to understand how ICT works, studies have suggested that ICL may involve implicit Bayesian inference [29] and a symbiotic relationship between text and patterns [17], and can behave similarly to explicit fine-tuning [8]. Our work focuses on the effect of demonstrations for ICL with large language models.

A.2 Retrieval Augmented Demonstrations

As summarized in Table 5, several previous works have explored retrieval techniques for identifying more informative demonstrations to boost in-context learning. KATE [14] discovers that semantically closer demonstrations outperform random ones for GPT-3 in-context learning. They employ language models trained on tasks like natural language inference and sentence textual similarity as semantic representations and utilize the kNN algorithm to search for demonstrations. EPR [24] develops a retriever based on language model signals to find superior demonstrations compared to off-the-shelf retrievers. Instead of using a separate retriever for each task, UPRISE [5] merges multiple training datasets into a retrieval corpus and trains a universal retriever for cross-domain tasks. PARC [21] employs a multilingual retrieval strategy to find demonstrations from high-resource tasks, thereby enhancing the performance of low-resource domain tasks. CEIL [30], instead of retrieving few-shot demonstrations independently, introduces an iterative retrieval method to identify both diverse and similar few-shot examples. While the aforementioned methods retrieve demonstrations from training data, [16] and [9] incorporate human feedback to create demonstrations and maintain a dynamic retrieval corpus. Z-ICL [15] generates pseudo demonstrations to enhance zero-shot in-context

Question	BM25 Demo	GTR Demo
Q: when does the new episodes of supernatural start? A: October 12, 2017	Q: when does the new episodes of ghost adventures start? A: June 16, 2018	Q: when does the next episode of supernatural come out? A: April 5, 2018
Kaj Birket-Smith (20 January 1893 – 28 October 1977) was a Danish philologist and anthropologist. He specialized in studying the habits and language of the Inuit and Eyak. He was a member of Knud Rasmussen's 1921 Thule expedition. In 1940, he became director of the Ethnographic Department of the National Museum of Denmark. question: Kaj Birket-Smith would have been a ripe old age of 128 if he were still alive today. Is it true, false, or neither? answer: false	Kaj Birket-Smith (20 January 1893 – 28 October 1977) was a Danish philologist and anthropologist. He specialized in studying the habits and language of the Inuit and Eyak. He was a member of Knud Rasmussen's 1921 Thule expedition. In 1940, he became director of the Ethnographic Department of the National Museum of Denmark. question: Kaj Birket-Smith was on the Thule expedition. Is it true, false, or neither? answer: true	Kaj Birket-Smith (20 January 1893 – 28 October 1977) was a Danish philologist and anthropologist. He specialized in studying the habits and language of the Inuit and Eyak. He was a member of Knud Rasmussen's 1921 Thule expedition. In 1940, he became director of the Ethnographic Department of the National Museum of Denmark. question: Kaj Birket-Smith was a very educated man about many different cultures and expressed love in his field of expertise. Is it true, false, or neither? answer: neither
Q: The original retail price of an appliance was 60 percent more than its wholesale cost. If the appliance was actually sold for 20 percent less than the original retail price, then it was sold for what percent more than its wholesale cost? Options: (A) 20(B) 28(C) 36(D) 40(E) 42Step-by-step reasoning process: wholesale cost = 100; original price = 100*1.6 = 160; actual price = 160*0.8 = 128. A: (B)	Q: A retail appliance store priced a video recorder at 20 percent above the wholesale cost of $200. If a store employee applied the 20 percent employee discount to the retail price to buy the recorder, how much did the employee pay for the recorder? Options: (A) $198 (B) $216 (C) $192 (D) $230 (E) $240 Step-by-step reasoning process: Wholesale cost of video recorder = 200 $ Video recorder was priced at 20 percent above 200 = 240 $ % discount given by store employee = 20 Emlpoyee paid = .8 * 240 = 192 $ A: (C)	Q: A retailer bought a machine at a wholesale price of $108 and later on sold it after a 10% discount of the retail price. If the retailer made a profit equivalent to 20% of the whole price, what is the retail price of the machine? Options: (A) 81 (B) 100 (C) 120 (D) 135 (E) 144 Step-by-step reasoning process: My solution: Wholesale Price= 108 Retail Price, be = x He provides 10 % discount on Retail price= x-10 x/100 This Retail price = 20 % profit on Wholesale price x-10 x/100 = 108+ 1/5(108) x=144; A: (E)

Table 6: Examples of retrieved demonstrations from NQ, ANLI(r3), and AQUQ.

performance. In contrast to the methods that retrieve explicit demonstrations, RETROPROMPT [4] transforms explicit demonstrations into implicit neural demonstrations represented by vectors. Rather than using a retriever, [22] applies a cross-attention reranker to re-rank documents retrieved by BM25. Our work distinguishes itself from previous research in that it integrates CoT prompting with a retriever and additionally examines the fine-tuned model based on instruction.

B Examples of Retrieved Demonstrations

In Table 6 and 7, we show the retrieved demonstrations given by BM25 and GTR.

Question	BM25 Demo	GTR Demo
Q: Lori wants to buy a $320.00 pair of shoes and a matching belt that is $32.00. Her part-time job pays her $8.00 an hour. How many hours will she have to work before she can make her purchase? Step-by-step reasoning process: b"She wants to buy a pair of shoes for $320.00 and a belt for $32.00 for a total of 320+32 = $«320+32=352.00»352.00 Her purchase will total $352.00 and she makes $8.00 at her part-time job so she'll have to work 352/8 = «352/8=44»44 hours A: 44	Q: Joanne makes $16.00 working at her main job for 8 hours a day. She has a part-time job, working an extra 2 hours a day where she makes $13.50 an hour. How much money does she make if she works this schedule 5 days a week? Step-by-step reasoning process: She works 8 hours a day at $16.00 an hour so she makes 8 * 16 = $128.00 a day. She works this job 5 days a week so she makes 128 * 5 = $640.00 in 5 days. She works 2 hours a day at $13.50 an hour so she makes 2 * 13.50 = $27.00 a day. She works this job 5 days a week so she makes 27 * 5 = $135.00. She makes $640 at her main job and $135 at her part - time job so all total she makes 640 + 135 = $775.00 in 5 days. A: 775	Q: Janice has been working part-time at a convenience store 5 days a week. She can earn $30 per day and can earn $15 more when she works a 2 hour overtime shift. If she works three overtime shifts this week, how much will she earn this week? Step-by-step reasoning process: Janice can earn $30 x 5 = $150 per week. She will earn $15 x 3 = $45 more if she works three overtime shifts. Therefore, Janice will earn $150 + $45 = $195 this week. A: 195
Q: If it socially acceptable to wear an icon depicting crucifixion? Step-by-step reasoning process: The crucifixion of Jesus is a common sign used by Catholics and Christian denominations.Many jewelry stores offer necklaces with the Crucifixion of Jesus Christ. A: yes	Q: Was the Donatello crucifix identified in 2020 life size? Step-by-step reasoning process: The crucifix discovered in the church of Sant'Angelo depicts an adult man. The crucifix discovered in the church of Sant'Angelo is 89 cm high. The crucifix discovered in the church of Sant'Angelo was identified as being a work of Donatello. The average height of an adult man has been at least 150 cm in historical times. A: no	Q: Did any cultures associate celery with death? Step-by-step reasoning process: Ancient Greeks used garlands of celery leafs to bury their dead. Ancient Greece was considered a culture. A: yes

Table 7: Examples of retrieved demonstrations from GSM8K and StrategyQA.