Probe then Retrieve and Reason: Distilling Probing and Reasoning Capabilities into Smaller Language Models

Yichun Zhao, Shuheng Zhou, Huijia Zhu
Ant Group
{zhaoyichun.zyc, shuheng.zsh, huijia.zhj}@antgroup.com

Abstract

Step-by-step reasoning methods, such as the Chain-of-Thought (CoT), have been demonstrated to be highly effective in harnessing the reasoning capabilities of Large Language Models (LLMs). Recent research efforts have sought to distill LLMs into Small Language Models (SLMs), with a significant focus on transferring the reasoning capabilities of LLMs to SLMs via CoT. However, the outcomes of CoT distillation are inadequate for knowledge-intensive reasoning tasks. This is because generating accurate rationales requires crucial factual knowledge, which SLMs struggle to retain due to their parameter constraints. We propose a retrieval-based CoT distillation framework, named Probe then Retrieve and Reason (PRR), which distills the question probing and reasoning capabilities from LLMs into SLMs. We train two complementary distilled SLMs, a probing model and a reasoning model, in tandem. When presented with a new question, the probing model first identifies the necessary knowledge to answer it, generating queries for retrieval. Subsequently, the reasoning model uses the retrieved knowledge to construct a step-by-step rationale for the answer. In knowledge-intensive reasoning tasks, such as StrategyQA and OpenbookQA, our distillation framework yields superior performance for SLMs compared to conventional methods, including simple CoT distillation and knowledge-augmented distillation using raw questions.

Keywords: large language models, chain-of-thought, distillation, knowledge retrieval

1. Introduction

Large Language Models (LLMs) have shown exceptional capabilities across a wide range of tasks in various domains through the method of in-context learning (Brown et al., 2020; Ouyang et al., 2022; OpenAI, 2023). Recent research suggests that increasing the number of parameters of LLMs can significantly enhance their knowledge encoding and reasoning capabilities (Wei et al., 2022a; Kaplan et al., 2020). Impressively, LLMs have excelled in domains demanding profound knowledge and reasoning, addressing significant challenges.

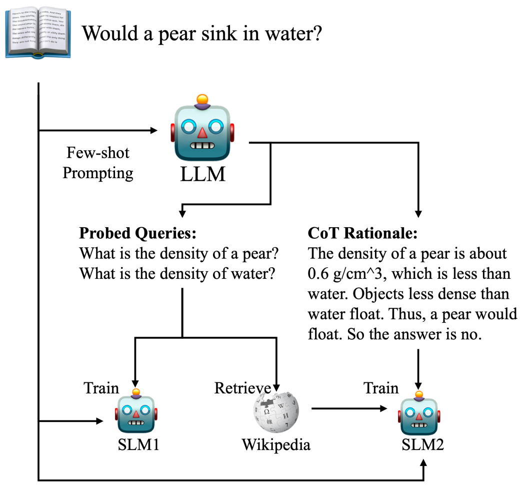
However, the real-world deployment of LLMs presents difficulties. Primarily, predictions from LLMs are computationally intensive. Furthermore, there are concerns regarding potential privacy breaches, as many commercially available LLMs (Brown et al., 2020; Ouyang et al., 2022; OpenAI, 2023) operate as opaque systems. These models typically restrict users to interfacing solely with the outputs, providing no access to or visibility of the underlying parameters.

In order to address deployment challenges, past studies (Ho et al., 2022; Li et al., 2022; Magister et al., 2022; Fu et al., 2023; Hsieh et al., 2023) have explored CoT distillation, aiming to transfer the reasoning capabilities of LLMs to Small Language Models (SLMs). These methods involve LLMs in creating high-quality rationales step by step, fine-tuning SLMs using these rationales. Such distillation techniques have significantly enhanced the capabilities of smaller models in tasks like arithmetic and symbolic reasoning (Cobbe et al., 2021; Wei et al., 2022b). However, the current distillation methods fall short in knowledge-intensive tasks because SLMs can't capture all necessary knowledge with their limited parameters. Consequently, there is a recognized need to incorporate task-specific knowledge during the distillation from LLMs to SLMs. (Kang et al., 2023) utilized a retriever (Robertson et al., 2009) to fetch relevant knowledge paragraphs from external databases like Wikipedia and fine-tune SLMs using both the raw questions and the retrieved paragraphs to generate rationales. However, the methodology in (Kang et al., 2023) has an inconsistency: it leverages LLM-generated rationales for knowledge retrieval during training but uses the raw questions for retrieval during inference, potentially compromising performance.

In this paper, we postulate that while SLMs lack the capacity for extensive knowledge storage, they can ascertain the requisite knowledge needed for question answering, which is a capability that can potentially be transferred from LLMs. Building on this insight, we propose a framework **Probe then Retrieve and Reason (PRR)** which separates the transfer of probing and reasoning from LLMs to two distinct SLMs, thereby improving retrieval and enriching the SLMs for knowledge-intensive tasks. Initially, an LLM is prompted to probe the raw question, dissecting it into related sub-queries that identify the necessary knowledge to answer the main question, and generating a rationale to answer it step by step. Subsequently, retrievers source paragraphs that correspond to the dissected sub-queries from

LREC-COLING 2024, pages 13026–13032
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

external knowledge bases. Finally, we fine-tune two SLMs based on the sub-queries and rationales obtained from LLMs with retrieved paragraphs. A high-level illustration of the process is provided in Figure 1.



Would a pear sink in water?
Few-shot Prompting
LLM
Probed Queries:
What is the density of a pear?
What is the density of water?
CoT Rationale:
The density of a pear is about 0.6 g/cm^3, which is less than water. Objects less dense than water float. Thus, a pear would float. So the answer is no.
Train
Retrieve
Train
SLM1
Wikipedia
SLM2

Figure 1: Illustration of the proposed framework PRR. SLM1 denotes the probing model and SLM2 denotes the reasoning model.

To demonstrate the effectiveness of **PRR**, we conducted an empirical evaluation using T5-base (Raffel et al., 2020) as a Small Language Model (SLM). Our results showed significant improvements compared to both basic CoT distillation and knowledge-augmented distillation using raw questions on multi-step factual QA datasets such as StrategyQA (Geva et al., 2021) and OpenbookQA (Mihaylov et al., 2018).

2. Related Works

Large Language Models Large Language Models (LLMs) have showcased impressive capabilities in a wide array of tasks. Their primary strength lies in the storage and utilization of knowledge for complex reasoning. For instance, LLMs such as GPT-3.5 (Ouyang et al., 2022) and GPT-4 (OpenAI, 2023) have delivered promising performances in various assessments. However, deploying LLMs in offline and privacy-conscious settings poses significant challenges, primarily because these models are often treated as black boxes accessible only via APIs, and are associated with substantial computational overheads. This calls for alternative approaches that utilize the full potential of LLMs in knowledge-intensive reasoning tasks.

CoT Distillation Recent studies have aimed to distill LLMs into SLMs (Ho et al., 2022; Li et al., 2022; Magister et al., 2022; Fu et al., 2023; Hsieh et al., 2023). A key focus of these efforts has been the transfer of specialized abilities like the Chain-of-Thought (CoT) paradigm from LLMs to SLMs, thereby enhancing their performance in arithmetic and complex reasoning tasks (Kojima et al., 2022; Wei et al., 2022b). (Shridhar et al., 2023) trained a combination of two small distilled models: a problem decomposer and a subproblem solver to reason better. However, prior research (Li et al., 2022; Ho et al., 2022) suggested that CoT distillation becomes less effective for knowledge-intensive reasoning tasks(Geva et al., 2021), as accurate rationales require a deep understanding of factual knowledge. (Kang et al., 2023) distilled SLMs using both the raw questions and the retrieved paragraphs to generate rationales while exhibits an inconsistency between the training and inference phases.

3. Methodology

3.1. Query and Rationale Generation from LLM

Based on prior research (Li et al., 2022), we harness the LLMs' capacity to identify knowledge for retrieval and to reason about their predictions in order to train SLMs. Let $D = \{(x_i, y_i)\}^N$ represent a dataset comprising N training instances, where x_i denotes a question and y_i denotes its corresponding answer. We also have a curated set of human-authored instances $E = \{(x_i^p, q_i^p, e_i^p, y_i^p)\}^M$, with $M \ll N$ (for our experiments, we set $M = 7$) based on prompts in (Li et al., 2022). In this set, q_i^p represents the probed queries that focus on the knowledge needed to answer the question x_i^p, while e_i^p is a free-text rationale elplaining why question x_i^p yields y_i^p as its answer. The set $(x_i^p, y_i^p)^M$ is a subset of D. Our primary objective is to optimally utilize the LLM, using E as reference demonstrations for in-context learning. This will enable the generation of queries q_i and rationale e_i for every (x_i, y_i) where $1 \leq i \leq N$. Subsequently, these LLM-generated queries and rationales can be used to augment the probing and reasoning capabilities of SLMs.

3.2. Training the Probing Model with Probed Queries

The first SLM which named the probing model, is trained with the objective of minimizing the query generation loss as follows:

$$\mathcal{L}_{probe} = \frac{1}{N} \sum_{i=1}^{N} l(f_1(x_i), q_i) \quad (1)$$

In this equation, f_1 represents the probing model, and l denotes the cross-entropy loss, which is com-

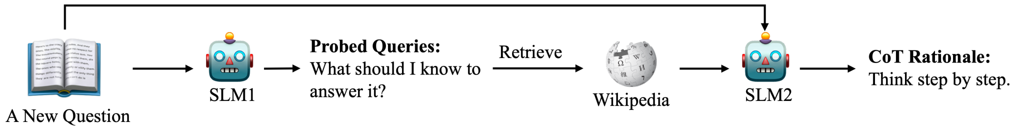


Figure 2: Illustration of the inferring process.

puted between the predicted tokens of the generated queries and the corresponding target tokens of the probed queries. The goal of this training process is to fine-tune the probing model to accurately generate queries that capture the necessary knowledge for answering the given questions.

3.3. Retrieving Knowledge with Probed Queries

We propose a method to retrieve paragraphs from an external Knowledge Base (KB) denoted as $\mathcal{P} = \{p_1, \cdots, p_k\}$. Retrieving the appropriate paragraph from $\mathcal{P}$ is crucial for training SLMs to produce high-quality rationales that, in turn, generate accurate answers to the raw questions. Following the approach of (Kang et al., 2023), we employ the sparse retriever BM25 (Robertson et al., 2009). To extract the most relevant knowledge for answering a given question, we use probed queries generated by the LLM during training or by the probing model at inference time. These queries facilitate the retrieval of a set of paragraphs given by $p_i^r = topk(\mathcal{R}(p^r|q_i;\mathcal{P}), k) \in \mathcal{P}$, where p_i^r represents the top k paragraphs with the highest relevance scores, as determined by the retriever $\mathcal{R}$ based on their relevance to the probed query q_i.

3.4. Training the Reasoning Model with Retrieved Knowledge and Rationales

To enhance the performance of the reasoning model, we leverage the raw question x_i along with the retrieved paragraphs p_i^r to generate an answer y_i and its corresponding rationale e_i for the question x_i. Two loss functions are given by:

$$\mathcal{L}_{answer} = \frac{1}{N}\sum_{i=1}^{N} l(f_2^{answer}(x_i, p_i^r), y_i) \qquad (2)$$

$$\mathcal{L}_{reason} = \frac{1}{N}\sum_{i=1}^{N} l(f_2^{reason}(x_i, p_i^r), e_i) \qquad (3)$$

where f_2 represents the reasoning model, $\mathcal{L}_{answer}$ denotes the label generation loss, while $\mathcal{L}_{reason}$ denotes the rationale generation loss.

The total loss function is a combination of both losses:

$$\mathcal{L} = \mathcal{L}_{answer} + \lambda\mathcal{L}_{reason} \qquad (4)$$

where λ is a hyperparameter that weights the importance of the rationale generation loss relative to the label generation loss. When λ is set to 0, the training process degenerates into a single-task fine-tuning that relies only on the raw classification labels.

3.5. Inferring with the Probing Model and the Reasoning Model

For inferring the dataset $\tilde{D} = \{(\tilde{x}_i, \tilde{y}_i)\}^I$, we first generate the probed query $\tilde{q}_i = f_1(\tilde{x}_i)$ with the probing model and then retrieve top-k paragraphs as $\tilde{p}_i^r = topk(\mathcal{R}(\tilde{p}^r|\tilde{q}_i;\mathcal{P})), k) \in \mathcal{P})$ from the external Knowledge Base. Finally, the answer $\tilde{y}_i = f_2(\tilde{x}_i, \tilde{p}_i^r)$ is generated by the reasoning model. An illustration of the inferring process is provided in Figure 2.

4. Experiments

4.1. Datasets and Metrics

We conducted experiments on the following two QA benchmark datasets:

StrategyQA is a dataset tailored for binary yes/no question-answering scenarios, which necessitates implicit multi-hop reasoning for inference (Geva et al., 2021). It includes 2,290 questions in the training set and 490 in the test set. Due to the inaccessibility of the official test set, we adopts an alternative approach using the split provided in their GitHub repository[1] where the original training set is partitioned randomly, allocating 90% for training and the residual 10% for the development set. We report results on their Github development set and utilize their Github training set for training, without utilizing rationales from their original annotations.

OpenbookQA is designed as a 4-way multiple-choice question-answering challenge, requiring extensive common knowledge as well as sophisticated multi-hop reasoning (Mihaylov et al., 2018). The dataset is divided into 4,957, 500, and 500 questions for the training, development, and test sets, respectively.

[1] https://github.com/eladsegal/strategyqa

	StrategyQA				OpenbookQA			
	WK	RQ	PQ	RQ+PQ	WK	RQ	PQ	RQ+PQ
ST	$58.08_{1.05}$	$56.10_{0.99}$	$58.10_{0.42}$	$56.25_{0.56}$	$55.07_{0.90}$	$54.23_{1.46}$	$\underline{66.20}_{0.29}$	$62.27_{1.55}$
MT-CoT	$\underline{59.82}_{0.24}$	$59.10_{1.29}$	$\mathbf{60.67}_{0.61}$	$57.20_{0.62}$	$62.27_{0.66}$	$60.87_{0.50}$	$\mathbf{68.67}_{0.77}$	$65.33_{0.66}$

Table 1: Accuracy comparison (%) of Single-Task fine-tuning (ST) and Multi-Task Chain-of-Thought fine-tuning (MT-CoT) utilizing knowledge retrieved with raw question (RQ), probed queries (PQ), both (RQ+PQ) or without knowledge (WK). Results are averaged over five runs with their standard deviation in the subscript. Best and second results for each dataset are bold and underlined.

To evaluate the question-answering performance on StrategyQA and OpenbookQA datasets, we employ the accuracy metric based on the final answer provided by the reasoning model. Each experiment is run 5 times with different random seeds, and we report the average accuracy score on the test set for reproducibility.

4.2. Implementation Details

In our experiments, we utilize the GPT-3.5-turbo model through the official OpenAI API[2] as the Large Language Model. For task-specific downstream applications, we employ the T5-base model (Raffel et al., 2020). Our framework is developed using PyTorch[3] and the Huggingface transformers library[4]. To implement BM25, we use the pyserini library[5] which provides a reproducible information retrieval framework. For CoT prompting, we follow the approach of (Li et al., 2022) and prepare our own examples tailored to new datasets. We empirically set the hyper-parameters and apply them consistently in all experiments. Specifically, the retrieving parameter k is set to 3, and the weight parameter λ in Eq. 4 is set to 0.5 to balance the label and rationale generation losses.

4.3. Main Results

The main results of our experiments on StrategyQA and OpenbookQA are shown in Table 1.

Overall, the Multi-Task Chain-of-Thought (MT-CoT) models, with the weight parameter λ set to 0.5 in Eq. 4, consistently outperform their Single-Task (ST) counterparts, where λ is set to 0, across a range of retrieval strategies. This result underscores the benefits of leveraging rationales from LLMs to augment the reasoning capabilities of SLMs.

The MT-CoT model demonstrates superior performance on the two datasets when utilizing knowledge retrieved through probed queries (PQ) as opposed to using knowledge retrieved through raw questions (RQ) or a combination of both (RQ+PQ). This underscores the potency of our Probe then Retrieve and Reason (PRR) framework. By distilling the probed queries and rationales generated by LLMs into SLMs and effectively integrating them, the quality of knowledge retrieval can be significantly improved, and the reasoning capabilities of the SLM can be more effectively harnessed.

Comparing the results between StrategyQA and OpenbookQA, we observe that the benefits from retrieval are more pronounced on OpenbookQA. This may be attributed to the fact that OpenbookQA is a more knowledge-intensive task. Additionally, when comparing different retrieval strategies, we find that using probed queries (PQ) yields greater improvements on OpenbookQA than on StrategyQA. We believe this is because the questions in StrategyQA are essentially standard queries without the explicit expression of the knowledge needed to answer them. In contrast, the questions in OpenbookQA are formatted as cloze tests. In this scenario, our probing model can not only identify the necessary knowledge but also transform the raw question into a refined query, thereby significantly enhancing the retrieval's effectiveness.

4.4. Analysis of the Necessity for Another Probing Model

In our study, we investigated an alternative training approach that avoids the use of an additional probing model. Instead, we integrated a query generation sub-task directly into the training of the reasoning model, effectively combining the two tasks into one. A comparative analysis of the model performance is presented in Table 2.

	StrategyQA	OpenbookQA
OM	$59.77_{0.28}$	$68.02_{0.34}$
TM	$\mathbf{60.67}_{0.61}$	$\mathbf{68.67}_{0.77}$

Table 2: Accuracy comparison (%) of one model (OM) and two model (TM) utilizing knowledge retrieved with probed queries (PQ).

According to the results, we found that using two separate SLMs sequentially yields better perfor-

[2] https://platform.openai.com/docs/models/gpt-3-5

[3] https://pytorch.org/

[4] https://github.com/huggingface/transformers

[5] https://github.com/castorini/pyserini

mance than using a single SLM. This suggests that when the model lacks sufficient parameters, the two tasks do not complement each other. It is more effective for the T5-base model to concentrate on a single task.

5. Conclusion

In this paper, we introduce the Probe then Retrieve and Reason (PRR) framework, which effectively distills probing and reasoning capabilities from Large Language Models (LLMs) into Small Language Models (SLMs). Our approach involves training two distilled SLMs in tandem: a probing model that probes into the question to generate queries identifying the necessary knowledge for retrieval, and a reasoning model that constructs a rationale to answer the question step by step, utilizing the retrieved knowledge. When applied to knowledge-intensive reasoning tasks such as StrategyQA and OpenbookQA, our PRR framework demonstrates superior performance for SLMs compared to traditional methods, including simple Chain-of-Thought (CoT) distillation and knowledge-augmented distillation with raw questions.

6. Acknowledge

This research work has been sponsored by Ant Group Security and Risk Management Fund.

7. Bibliographical 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.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. 2023. Specializing smaller language models towards multi-step reasoning. *arXiv preprint arXiv:2301.12726*.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2022. Large language models are reasoning teachers. *arXiv preprint arXiv:2212.10071*.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *arXiv preprint arXiv:2305.02301*.

Minki Kang, Seanie Lee, Jinheon Baek, Kenji Kawaguchi, and Sung Ju Hwang. 2023. Knowledge-augmented reasoning distillation for small language models in knowledge-intensive tasks. *arXiv preprint arXiv:2305.18395*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, et al. 2022. Explanations from large language models make small reasoners better. *arXiv preprint arXiv:2210.06726*.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2022. Teaching small language models to reason. *arXiv preprint arXiv:2212.08410*.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

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

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. 2023. Distilling reasoning capabilities into smaller language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7059–7073.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

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

8. Language Resource References

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391.