Teaching Small Language Models to Reason for Knowledge-Intensive Multi-Hop Question Answering

Xiang Li[1,2], **Shizhu He**[1,2*], **Fangyu Lei**[1,2], **Jun Yang**[3], **Tianhuang Su**[3], **Kang Liu**[1,2,4], **Jun Zhao**[1,2]
[1]The Laboratory of Cognition and Decision Intelligence for Complex Systems, Institute of Automation, Chinese Academy of Sciences
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences
[3] Guangdong OPPO Mobile Telecommunications Corp.,Ltd.
[4] Shanghai Artificial Intelligence Laboratory
{lixiang2022, leifangyu2022}@ia.ac.cn {shizhu.he, kliu, jzhao}@nlpr.ia.ac.cn
{yangjun2, sutianhuang}@oppo.com

Abstract

Large Language Models (LLMs) can teach small language models (SLMs) to solve complex reasoning tasks (e.g., mathematical question answering) by Chain-of-thought Distillation (CoTD). Specifically, CoTD fine-tunes SLMs by utilizing rationales generated from LLMs such as ChatGPT. However, CoTD has certain limitations that make it unsuitable for knowledge-intensive multi-hop question answering: 1) SLMs have a very limited capacity in memorizing required knowledge compared to LLMs. 2) SLMs do not possess the same powerful integrated abilities in question understanding and knowledge reasoning as LLMs. To address the above limitations, we introduce Decompose-and-Response Distillation (D&R Distillation), which distills two student models, namely *Decomposer* and *Responser* separately. The two models solve a knowledge-intensive multi-hop question through an interactive process of asking and answering subquestions. Our method offers two advantages: 1) SLMs have the capability to access external knowledge to address subquestions, which provides more comprehensive knowledge for multi-hop questions. 2) By employing simpler subquestions instead of complex CoT reasoning, SLMs effectively mitigate task complexity and decrease data prerequisites. Experimental results on three knowledge-intensive multi-hop question answering datasets demonstrate that D&R Distillation can surpass previous CoTD methods, even with much less training data[1].

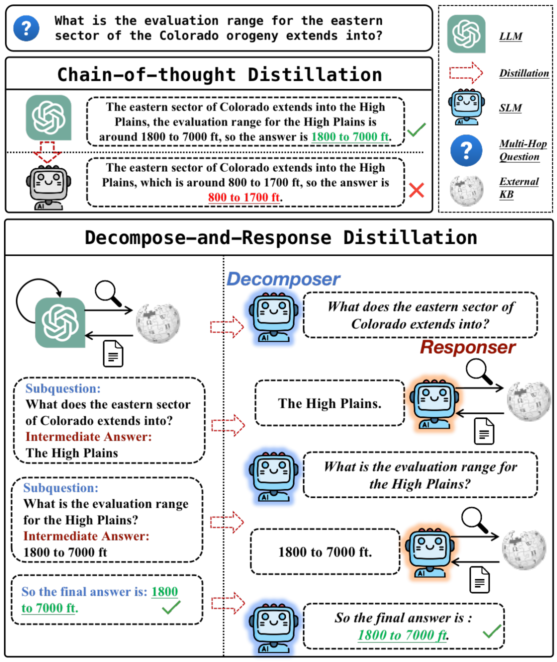


Figure 1: A comparison of D&R Distillation (ours) and CoTD (Ho et al., 2023). CoTD teaches one SLM to output all intermediate reasoning steps and the final answer at once, struggling on knowledge-intensive multi-hop questions. D&R Distillation teaches two SLMs to interact by asking and answering subquestions, leading them to collectively reach the final answer.

1 Introduction

Large language models are capable of answering complex questions (e.g., mathematical questions) by generating step-by-step natural language reasoning paths, namely Chains-of-thoughts (CoTs) (Wei et al., 2022). However, the ability to solve complex reasoning tasks through CoT prompting is considered an emergence that appears in very large models with at least tens of billions of parameters (Wei et al., 2022), such as PaLM of 540B (Chowdhery et al., 2022), GPT-3 of 175B (Brown et al., 2020), and LLaMA of 70B (Touvron et al., 2023).

Recent works have proposed to transfer the rea-

*Corresponding author
[1]Our code will be available at `https://github.com/Xiang-Li-oss/D-R-Distillation`

Findings of the Association for Computational Linguistics: ACL 2024, pages 7804–7816
August 11-16, 2024 ©2024 Association for Computational Linguistics

soning ability of large models to small language models (SLMs) through Chain-of-thought Distillation (CoTD) (Ho et al., 2023; Magister et al., 2023; Li et al., 2023a). Specifically, as shown in the upper part of Figure 1, they leverage the LLM (e.g., ChatGPT) to generate high-quality rationales and fine-tune a SLM with rationale-augmented question-answer pairs. CoTD has successfully enhanced SLMs' reasoning ability on many reasoning tasks, such as arithmetic reasoning (Cobbe et al., 2021), commonsense reasoning (Talmor et al., 2019), and symbolic reasoning (Wei et al., 2022).

However, previous CoTD works did not effectively address knowledge-intensive reasoning tasks such as multi-hop question answering (Petroni et al., 2021; Trivedi et al., 2023). Unlike arithmetic reasoning and commonsense reasoning, knowledge-intensive reasoning tasks pose greater challenges due to their requirement for both background knowledge and the ability to perform multi-step reasoning. CoTD has two limitations that render it unsuitable for teaching SLM to reason over knowledge-intensive multi-hop question answering.

1) **Knowledge Memorization Gap between LLMs and SLMs**. Unlike LLMs, which store vast amounts of knowledge within their parameters, SLMs are limited in their capacity to memorize the necessary knowledge to solve the tasks due to their small number of parameters. Besides, simply augmenting SLM with a *one-step* retrieval-augmentation strategy (Kang et al., 2023; Zhang et al., 2023) is also suboptimal for multi-hop questions. For such questions, relevant knowledge often needs to be retrieved after intermediate reasoning has concluded, as it may not be explicitly mentioned in the question. For example, consider the question illustrated in Figure 1, one must first infer that `the eastern sector of Colorado extends into the High Plains`, and then perform further retrieval to obtain evidence pointing to the `evaluation range`.

2) **Difficulty in Distilling Integrated Subtasks**. In contrast to arithmetic reasoning, which typically involves applying predefined formulas or algorithms, or commonsense reasoning, which relies on general knowledge and intuition. Solving a knowledge-intensive multi-hop question via chain-of-thought reasoning potentially involves a collection of multiple subtasks, including complex question decomposition, knowledge association, and knowledge reasoning (Zheng et al., 2023). However, it is highly challenging for an individual SLM to simultaneously acquire all these integrated capabilities, which leads to the CoTD methods requiring more training data and being inefficient.

To address the aforementioned limitations, motivated by question decomposition for answering complex questions (Han et al., 2023; Press et al., 2023), we propose a novel method to teach SLMs to reason for knowledge-intensive multi-hop questions, namely Deompose-and-Response Distillation (D&R Distillation, as shown in Figure 1). Specifically, we propose to prompt LLM in a *Self-Ask-Self-Ans* strategy by iteratively asking subquestions and responding with intermediate answers. Then we separately distill two student models, namely *Decomposer* and *Responser*. The *Deomposer* is responsible for asking subquestions and determining the final answer based on current interaction history. The *Responser* is responsible for answering subquestions by leveraging relevant background knowledge obtained from an external knowledge base. By formatting the reasoning process as a sequence of generating subquestions and intermediate answers, these two student models effectively address knowledge-intensive multi-hop questions within an interactive framework.

Compared with previous Chain-of-thought Distillation methods, our method offers two notable advantages: 1) By reasoning in an interactive manner, our method allows student models to utilize external knowledge with each retrieval focusing on a subquestion. Compared to previous works relying solely on parameter knowledge or *one-step* retrieval augmentation (Ho et al., 2023; Kang et al., 2023), our method provides a more comprehensive collection of relevant knowledge required to answer multi-hop questions. 2) We transform the process of solving a reasoning question into two interrelated and decoupled subtasks: decomposing the complex question and solving a series of simpler subquestions. D&R Distillation effectively reduces the overall task difficulty while significantly reducing the amount of data required for distillation.

We evaluate the effectiveness of our method on three knowledge-intensive multi-hop question answering datasets: HotpotQA, StrategyQA, and 2WikiMultiHopQA. Experimental results demonstrate that D&R distillation significantly improves the knowledge-intensive reasoning ability of SLMs with approximately 1/10 of the full training data. Notably, our method with two 220M SLMs (T5-base) outperforms Chain-of-thought Prompting with an 11B (50 times larger) LLM (Flan-T5-XXL) on HotpotQA and 2WikiMultiHopQA.

2 Related Work

Chain-of-Thought prompting (Wei et al., 2022) significantly enhances the reasoning capacities of large language models by augmenting few-shot examples with detailed reasoning steps. Recent works have further refined CoT through verification (Li et al., 2023b), question decomposition (Zhou et al., 2023), and path sampling (Wang et al., 2023; Yao et al., 2023). However, these aforementioned studies primarily concentrate on enhancing the reasoning capabilities of LLMs, neglecting the necessity to improve the reasoning abilities of smaller language models (<1B).

Chain-of-thought Distillation have been proposed to distill the CoT reasoning ability of LLMs into SLMs (Ho et al., 2023; Fu et al., 2023; Magister et al., 2023; Hsieh et al., 2023), because the CoT reasoning ability is considered as an emergent ability which enables LLM to generate intermediate reasoning steps with CoT prompting (Wei et al., 2022) (e.g. Let's think step by step). To augment Chain-of-thought Distillation (CoTD) with external knowledge, (Kang et al., 2023) augment SLMs with documents retrieved by a *one-step* retriever from the external knowledge base. However, CoTD is less effective for knowledge-intensive multi-hop question answering tasks (Petroni et al., 2021), where both factual knowledge and multi-hop reasoning are important to generate accurate rationale. In this paper, we propose to distill two student models and solve a knowledge-intensive multi-hop question by facilitating an interactive process of asking and answering subquestions between the two student models.

Question Decomposition (Kalyanpur et al., 2012; Patel et al., 2022) has long been a crucial technique for understanding and solving complex questions. Recent works also utilize question decomposition to improve the reasoning ability of LLMs. (Zhou et al., 2023) enhances the CoT reasoning ability of LLMs by decomposing questions into subquestions and sequentially solving subquestions. (Press et al., 2023) explicitly asks LLM itself follow-up subquestions before answering the original question and answers subquestions with an external search engine. (Shridhar et al., 2023) learns a semantic decomposition of the original question into a sequence of subquestions and uses it to train two models designated for question decomposition and resolution. Unlike the aforementioned works, we focus on teaching small language models to reasoning for knowledge-intensive multi-hop questions with LLM generations. We achieve this by distilling two student models to interactively ask and answer subquestions.

3 Method

In this section, we provide a detailed description of our method. As illustrated in Figure 2, D&R Distillation can be divided into three stages:

1) *Self-Ask-Self-Ans* Prompting: We prompt a very large language model (e.g., ChatGPT) to generate D&R Distillation samples, preparing datasets for training student models.

2) *Decomposer* and *Responser* Training: We distill two student models (e.g., T5) with D&R Distillation samples obtained by stage 1).

3) *Decomposer* and *Responser* Interaction: The *Decomposer* and the *Responser* address a knowledge-intensive multi-hop question through an interactive process of generating subquestions and obtaining intermediate answers.

3.1 *Self-Ask-Self-Ans* Prompting

In this stage, a teacher model (LLM) is prompted with *Self-Ask-Self-Ans* prompting to generate D&R Distillation samples[2]. Specifically, the teacher model solves a knowledge-intensive multi-hop question by iteratively asking itself subquestions and providing intermediate answers. Consider a standard sample S_i consisting of a question q_i and its golden answer a_i. The teacher model serves as a *Decomposer* and a *Responser* alternatively. At the k-th step, when serving as a *Decomposer*, the teacher model decide to continue asking a subquestion s_i^k or predicting the final answer a_i^k based on interaction history:

$$H =< q_i, s_i^1, r_i^1, ..., s_i^{k-1}, r_i^{k-1} >$$

where s_i^t and r_i^t are the subquestion and the intermediate answer of the t-th step. When serving as a *Responser*, the teacher model answers the subquestion s_i^k proposed before with retrieved passages:

$$P_i^k = topK(R(p|s_i^k; D), K)$$

$$r_i^k = LLM(P_i^k, s_i^k)$$

[2]Prompting examples for the teacher model can be found in Appendix B

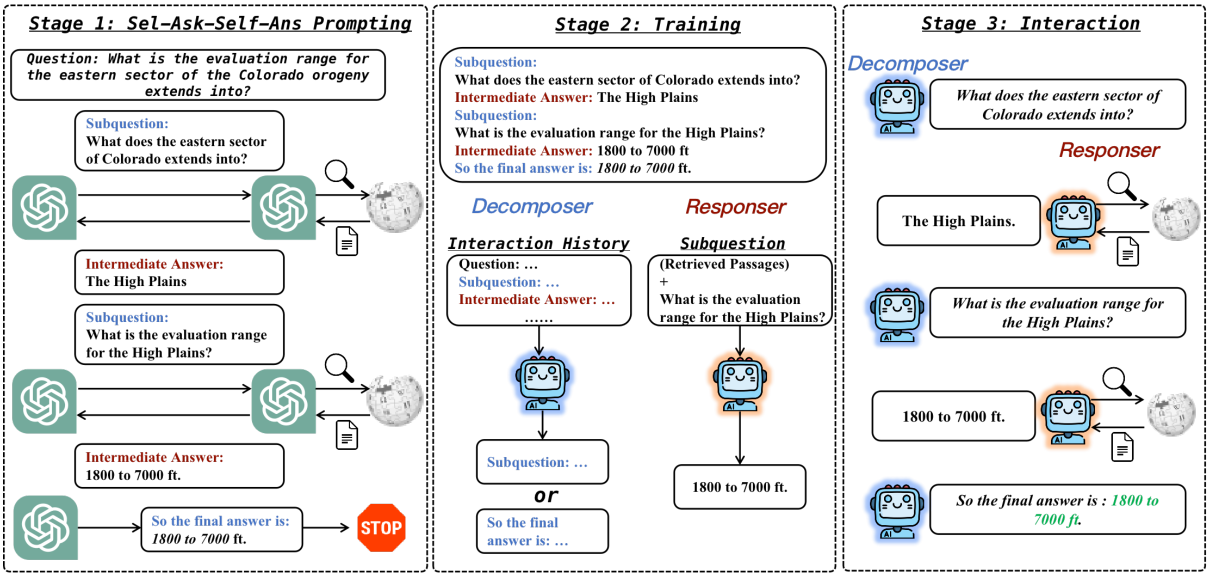


Figure 2: Overview of our proposed D&R Distillation method. **Stage 1:** A large language model is prompted to solve a knowledge-intensive multi-hop question by generating a series of subquestions and intermediate answers. This interaction process is used to compose D&R Distillation samples. **Stage 2**: D&R Distillation samples are used to finetune two student models, the *Decomposer* and the *Responser*. The *Decomposer* is responsible for asking subquestions or determining the final answer based on current interaction history and the *Responser* is responsible for answering subquestions with retrieved knowledge. **Stage 3**: The *Decomposer* and the *Responser* solve a knowledge-intensive multi-hop question in an interactive process.

where R is a retriever and D is a knowledge base (e.g., Wikipedia). Once the teacher model decide to predict the final answer a_i^k, we obtain a D&R Distillation sample $(q_i, s_i^1, r_i^1, ..., s_i^{k-1}, r_i^{k-1}, a_i^k)$. Moreover, to control the quality of generated samples, we filter generated D&R Distillation samples by comparing the final prediction a_i^k of the teacher model with the ground truth a_i. More detailed filter criteria can be found in Appendix A.

3.2 *Decomposer* and *Responser* Training

After acquiring D&R Distillation samples, we use them to fine-tune two small student models, namely the *Decomposer* p_θ^d and the *Responser* p_ϕ^r with trainable parameters θ and ϕ respectively. Specifically, consider a D&R Distillation sample $(q_i, s_i^1, r_i^1, ..., s_i^{k-1}, r_i^{k-1}, a_i^k)$, for the *Decomposer*, we minimize the negative log-likelihood of the sequence of subquestions $s_i^j (j = 1, 2, ..., k-1)$ and the final answer a_i^k:

$$L_D(\theta) = -\sum_{i=1}^{N}\sum_{j=1}^{k} \log p_\theta^d(o_i^j|H) \qquad (1)$$
$$(o_i^j = a_i^j \; if \; j = k \; else \; s_i^j)$$

where H represents the interaction history before j-th step:

$$H =< q_i, s_i^1, r_i^1, ..., s_i^{j-1}, r_i^{j-1} >$$

For the *Responser*, we minimize the negative log-likelihood of the sequence of intermediate answer r_i^j with augmented external knowledge:

$$P_i^j = topK(R(p|s_i^j; D), K)$$
$$L_R(\phi) = -\sum_{i=1}^{N}\sum_{j=1}^{k} \log p_\phi^r(r_i^j|s_i^j, P_i^j) \qquad (2)$$

where R is the same retriever in 3.1.

3.3 *Decomposer* and *Responser* Interaction

This section describes the behavior of two student models in the inference stage. After the aforementioned two stages, the *Decomposer* and the *Responser* work interactively to jointly solve a knowledge-intensive multi-hop question. As shown in Algorithm 1, we initiate with feeding the initial question to the *Decomposer*, at the j-th step, the *Decomposer* decides whether to ask another subquestion or predict the final answer based on current interaction history H. If the generation of the *Decomposer* is another subquestion, then the *Responser* retrieves related knowledge from a

Algorithm 1 Inference of D&R Distillation

Initialization: $H = \{q_i\}$, MAXSTEP $\leftarrow T$, $j \leftarrow 0$, p_θ^d, p_ϕ^r, R, D, K
repeat
  $o_i^j = argmax_o p_\theta^d(o|H)$
  if o_i^j is subquestion **then**
    $P_i^j = topK(R(p|s_i^j; D), K)$
    $r_i^j = argmax_r p_\phi^r(r|o_i^j, P_i^j)$
    H.append(o_i^j, r_i^j)
  end if
  if o_i^j is final answer **then**
    break
  end if
  $j \leftarrow j + 1$
until j =MAXSTEP
Output: final answer o_i^j

knowledge base and generates a response to the subquestion. Otherwise, if the generation of the *Decomposer* is the final answer, the interaction terminates and returns the final answer.

4 Experiments

4.1 Datasets

We evaluate our method on three knowledge-intensive multi-hop question answering datasets in the open-domain setting: **HotpotQA** (Yang et al., 2018), **2WikiMultiHopQA** (Ho et al., 2020), and **StrategyQA** (Geva et al., 2021). In contrast to previous works (Ho et al., 2023) of fine-tuning with the entire training set, we only fine-tune our model with 8800 instances (1/10 of the full training data) for HotpotQA, 16000 instances (1/10 of the full training data) for 2WikiMultiHopQA, and 1200 (1/2 of the full training data) instances for StrategyQA, eliminating the need for generating a large number of rationales with LLMs.

4.2 Teacher and Student Models

For teacher models, we use GPT3.5 (Brown et al., 2020) provided by the OpenAI API. Unless otherwise stated, we use `gpt3.5-turbo-instruct` as the teacher model. For student models, we adopt T5-{Small, Base, Large} (Raffel et al., 2020).

4.3 Baseline Methods

We provide a comparison of D&R Distillation (ours) with four baseline methods: **Fine-tuning** directly fine-tunes a student model to generate an answer given only a question (Petroni et al., 2021). **CoT Distillation** finetunes a student model with LLM-generated rationales, which is a typical approach for enhancing the reasoning capabilities of SLMs (Ho et al., 2023). The above baselines measure the capability of a small language model to solve knowledge-intensive multi-hop question answering relying only on parameter knowledge but without any external knowledge.

Retrieval-Augmented **Fine-tuning** appends retrieved passages along with the question at both training and inference time (Petroni et al., 2021). ***Retrieval-augmented*** **CoT Distillation** augments CoT Distillation with retrieved passages for both teacher and student models (Kang et al., 2023). The above two baselines help us to investigate the impact of incorporating external knowledge.

4.4 Implementation Details

We fine-tune student models for a maximum of 20 epochs with Pytorch-Lightning library[3], setting the batch size at 16 and the learning rate at $3e-4$.

For *Retrieval-augmented* methods, we use Wikipedia as the external knowledge base. For a fair comparison, we use the sparse retrieval method BM25 as the retriever provided by Pyserini library [4] for all baseline methods and our method. See Appendix A for more detail.

4.5 Experimental Results

In this section, we present the knowledge-intensive reasoning performance of our D&R Distillation. We compare our method with various baselines across different model sizes.

As shown in Table 1, the improvement of Chain-of-thought Distillation (CoT Distillation) compared to Fine-tuning is quite limited, and in some cases, even a performance decline has been observed. For example, T5-base exhibits a mere 0.9% (32.5%-31.6%) increase in Answer F1 on 2WikiMultiHopQA whereas it encounters a 0.4% (19.3%-19.7%) drop in Answer F1 on HotpotQA. This phenomenon can be highly attributed to the lack of background knowledge. Although CoT Distillation trains SLMs with augmentation of intermediate reasoning steps, it remains a challenge for SLMs to effectively reason without the necessary background knowledge.

The application of retrieval augmentation benefits both Fine-tuning and CoT Distillation. For example, the utilization of retrieval augmentation

[3] https://lightning.ai
[4] https://github.com/castorini/pyserini

Method	Params	*Data Usage*	HotpotQA		2WikiMultiHopQA		StrategyQA
			Answer EM	Answer F1	Answer EM	Answer F1	Answer Acc
			Teacher: GPT3.5 (`gpt3.5-turbo-instruct`)				
Few-shot-CoT	175B	-	35.6	49.2	36.5	43.9	66.4
			Student: T5 (`small`, `base`, `large`)				
Fine-tuning (Petroni et al., 2021)	60M 220M 700M	*All*	12.6 13.1 14.7	19.3 19.7 22.1	26.2 27.8 28.9	30.3 31.6 32.9	51.5 52.3 56.3
Retrieval-augmented Fine-tuning (Petroni et al., 2021)	60M 220M 700M	*All*	14.6 (+2.0) 15.2 (+2.1) 17.3 (+2.6)	21.5 (+2.2) 22.1 (+2.4) 23.8 (+1.7)	27.4 (+1.2) 29.1 (+1.3) 31.2 (+2.3)	32.4 (+2.1) 33.6 (+2.0) 35.4 (+2.5)	51.1 (-0.4) 52.1 (-0.2) 58.8 (+2.5)
CoT Distillation (Ho et al., 2023)	60M 220M 700M	*All*	12.2 (-0.4) 12.5 (-0.6) 16.9 (+2.2)	19.1 (-0.2) 19.3 (-0.4) 23 (+0.9)	26.8 (+0.6) 28.3 (+0.5) 30.6 (+1.7)	31.5 (+1.2) 32.5 (+0.9) 33.6 (+0.7)	52.8 (+1.3) 55.3 (+3.0) 64.4 (+8.1)
Retrieval-augmented CoT Distillation (Kang et al., 2023)	60M 220M 700M	*All*	14.5 (+1.9) 14.7 (+1.6) 18.2 (+3.5)	21.6 (+2.3) 22.2 (+2.5) 25.5 (+3.4)	28.3 (+2.1) 30.1 (+2.3) 32.0 (+3.1)	32.7 (+2.4) 34.6 (+3.0) 35.8 (+2.9)	53.3 (+1.8) 56.6 (+4.3) **65.0 (+8.7)**
D&R Distillation *(ours)*	60M 220M 700M	*1/10 or 1/2*	**18.2 (+5.6)** **19.9 (+6.8)** **21.7 (+7.0)**	**26.1 (+6.8)** **27.9 (+8.2)** **30.4 (+8.3)**	**29.5 (+3.3)** **32.5(+4.7)** **34.7 (+5.8)**	**33.7 (+3.4)** **37.0 (+5.4)** **39.4 (+6.5)**	**55.0 (+3.5)** **59.0 (+6.7)** 63.3 (+7.0)

Table 1: **D&R Distillation Performance.** Answer EM/F1/Acc (%) of student models on three knowledge-intensive multi-hop question answering datasets with D&R Distillation and baseline methods. (+/-) refers to the performance gain/drop compared to the Fine-tuning baseline. For the larger-scale HotpotQA and 2WikiMultiHopQA datasets, D&R Distillation only uses **1/10** of the full training data, and for the smaller-scale StrategyQA dataset, D&R Distillation only uses **1/2** of the full training data.

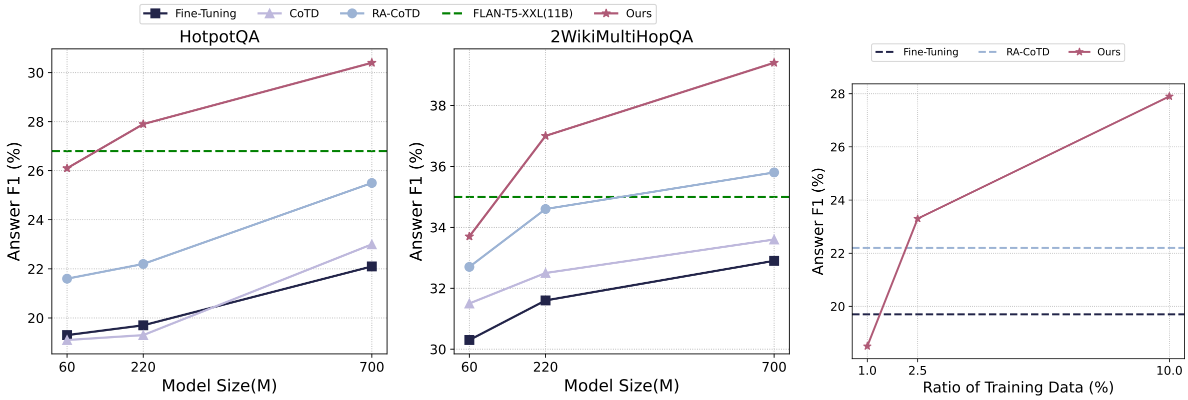
leads to a noteworthy improvement in the performance of T5-base. It enhances the Answer F1 of HotpotQA from 19.3% to 22.2% and increases the Answer accuracy of StrategyQA from 55.3% to 56.6%. However, augmenting CoT Distillation with a one-step retriever alone can not achieve comparable results to our method except for the StrategyQA dataset with T5-large. We attribute this discrepancy to the nature of the StrategyQA dataset, which consists of relatively easier yes/no questions. Therefore, it becomes easier for a model to find shortcuts to reach the final answer.

In contrast, D&R Distillation improves the knowledge-intensive reasoning ability of SLMs by a large margin and surpasses all baseline methods with student models of different sizes. Moreover, the performance gap between D&R Distillation and Fine-tuning baseline enlarges as the number of parameters of the student model increases. With T5-large, D&R Distillation achieves an Answer F1 gain of 8.3% and 6.5% over Fine-tuning on HotpotQA and 2WikiHotpotQA respectively.

Furthermore, it is noteworthy that our approach is trained using a significantly smaller fraction of the data compared to the baseline methods. For the larger-scale HotpotQA and 2WikiMultiHopQA datasets, we utilize only 1/10 of the training data, while for the smaller-scale StrategyQA dataset, we use only 1/2 of the training data. The above findings highlight the significant advantages of our method in terms of both performance and efficiency. Unlike existing (*Retrieval-augmented*) CoT Distillation methods, which heavily rely on extensive CoT annotations but struggle to effectively enhance the model's knowledge-intensive reasoning capabilities, our approach achieves superior performance, despite utilizing only a small fraction of data.

4.6 Analysis

Efficiency on Model Size and Training Data To validate the efficiency of our D&R Distillation method in terms of model size and training data, we measure the Answer F1 on HotpotQA and 2WikiMultiHopQA varying model parameters and the Answer F1 on HotpotQA varying the number of training data. As shown in Figure 3a, D&R Distillation consistently outperforms the CoTD and RA-CoTD baselines varying different model sizes with



(a) Efficiency on model size

(b) Efficiency on training data

Figure 3: **(a) Efficiency on model size and (b) training data.** On HotpotQA and 2WikiMultiHopQA, we compare D&R Distillation against CoT Distillation (CoTD) and *Retrieval-augmented* CoT Distillation (RA-CoTD) baselines, by varying the number of parameters, including the few-shot in-context learning performance of Flan-T5-XXL (11B). On HotpotQA, we compare D&R Distillation varying the number of training data with Fine-tuning and RA-CoTD baseline with full training data.

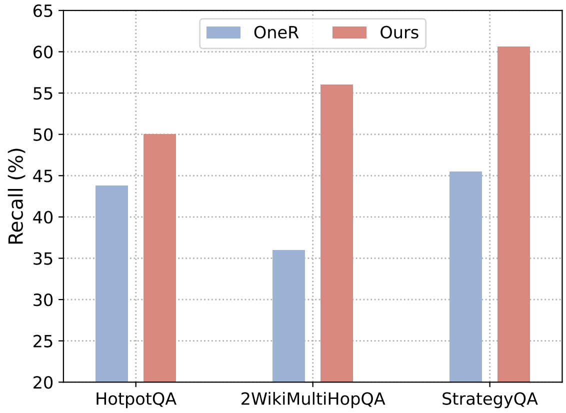


Figure 4: Retrieval Recall for one-step retriever (OneR) adopted in *retrieval-augmented* baseline methods and our D&R Distillation method. D&R Distillation demonstrates a significant performance improvement compared to OneR.

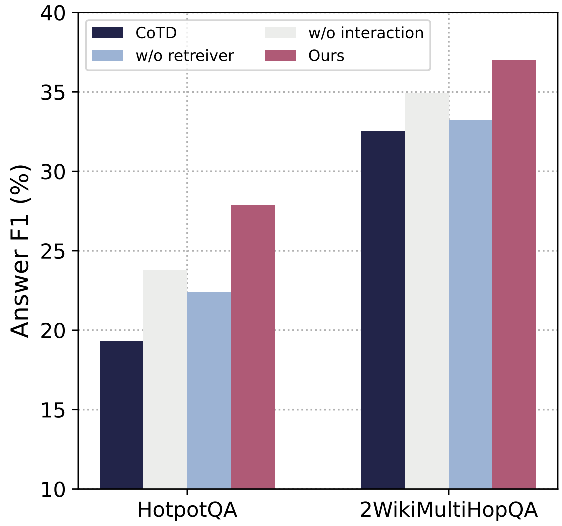


Figure 5: Ablation study on the multi-step retrieval and interactive process of D&R Distillation denoted as **w/o retriever** and **w/o interaction** respectively. The absence of either one of them will harm the answer F1 for answering knowledge-intensive multi-hop questions.

only 1/10 of the entire training dataset. Notably, on the HotpotQA dataset, D&R Distillation with two 60M student models achieves higher Answer F1 than CoTD with a 700M student model, whether enhanced with *Retrieval augmentation*. Moreover, D&R Distillation with two 220M student models outperforms the 11B LLM (FLAN-T5-XXL) in-context learning baseline. This observation shows a significant practical advantage of our approach in resource-restricted settings since the SLM with D&R Distillation requires significantly less computational cost yet it outperforms the LLM.

As shown in Figure 3b, the proposed D&R Distillation method can successfully transfer the knowledge-intensive reasoning ability, using only a small number of training data. Specifically, with 10% of the training data, D&R Distillation significantly outperforms the same model distilled by RA-CoTD on the full data by 5.7% on answer F1. With 2.5% of the training data, D&R Distillation still outperforms RA-CoTD on the full data. This indicates that we can potentially further increase the effectiveness of D&R Distillation with more training data.

Evaluation on Knowledge Acquisition To evaluate the effectiveness of D&R Distillation on retrieving relevant knowledge for knowledge-intensive multi-hop question answering, we compare the retrieval recall of our method with one-step retriever (OneR) adopted in *Retrieval-augmented*

Method	Retrieved Passages	Outputs
Question: In what year was the novel that Louren Mutarelli based "Nina" on first published? (1866)		
RA-CoTD	(...)Lourenço Mutarelli has also created plays, books, and all the illustrations and animated sequences of the 2004 motion picture "Nina", which was based on the novel "Crime and Punishment" by Dostoyevsky(...)	The novel that Lourenço Mutarelli based "Nina" on, titled "Crime and Punishment", was first published in 2002.
Ours	(...)Lourenço Mutarelli has also created plays, books, and all the illustrations and animated sequences of the 2004 motion picture "Nina", which was based on the novel "Crime and Punishment" by Dostoyevsky(...) (...)Crime and Punishment (...) It was first published in the literary journal "The Russian Messenger" in twelve monthly installments during 1866.(...)	What is the name of the novel that Louren Mutarelli based "Nina" on? Crime and Punishment. When was Crime and Punishment first published? 1866. So the final answer is: 1866.
Question: Who is older, Annie Morton or Terry Richardson? (Terry Richardson)		
RA-CoTD	Annie Morton (born October 8, 1970) is an American model(...) Terrence Richardson (born August 14, 1965) is an American fashion(...)	Annie Morton was born on October 8, 1970, while Terry Richardson was born on August 14, 1965. Therefore, Annie Morton is older.
Ours	Annie Morton (born October 8, 1970) is an American model(...) Terrence Richardson (born August 14, 1965) is an American fashion(...)	When was Annie Morton born? Annie Morton was born on October 8, 1970. When was Terry Richardson born? Terry Richardson was born on August 14, 1965. So the final answer is: Terry Richardson

Table 2: **Case Study** of D&R Distillation (Ours), compared with *Retrieval-augmented* Distillation (RA-CoTD) on HotpotQA with T5-base. The gold answer is in blue and the correct/wrong answer is marked as green/red We highlight supporting facts in the passages as yellow.

baseline methods. As shown in Figure 4, our method achieved significantly higher recall compared to OneR. Particularly, D&R Distillation demonstrates a remarkable 20.6% superiority in recall over OneR on the 2WikiMultiHopQA dataset. This indicates that by decomposing and retrieving based on subquestions iteratively, D&R Distillation obtains a more sufficient set of knowledge to answer knowledge-intensive multi-hop questions.

Ablation Study We conduct an ablation study to demonstrate the effectiveness of two designs in our method: 1) incorporating multi-step retrieval based on subquestions and 2) interaction process between *Decomposer* and *Responser*. For 1), we disable the retriever and do not provide retrieved passages for *Responser*, denoted as **w/o retriever**. For 2) we train *Decomposer* to output all subquestions at once and train the *Responser* to output all intermediate answers, as well as the final answer at once, denoted as **w/o interaction**. We then compare the Answer F1 of the two ablation settings with our original design and the CoT Distillation (CoTD) baseline. As shown in Figure 5, both of these designs are crucial for our method, as the absence of either one would result in performance degradation. On the other hand, the performance without either of these designs still surpasses that of CoTD, demonstrating their strength. The performance decline becomes even more pronounced when the retriever is removed (w/o retriever), further confirming the crucial role of background knowledge for knowledge-intensive multi-hop reasoning.

Case Study In Table 2, we provide two examples from the HotpotQA dataset comparing the output generated by our D&R Distillation against the rationale by the baseline method *Retrieval-augmented* CoT Distillation (RA-CoTD). For the first question, RA-CoTD fails to retrieve a passage about `Crime and Punishment`, as a result, it mistakenly generates the hallucination that `"Crime and Punishment" was first published in 2002`. For the second question, RA-CoTD successfully retrieved the necessary knowledge for answering the question, however, it fails to perform correct reasoning by mistakenly assuming that `Annie Morton (born in 1970) is older than Terry Richardson (born in 1965)`.

In contrast, D&R Distillation successfully retrieves a passage about `Crime and Punishment` by first generating subquestion `When was Crime and Punishment first published` and retrieving based on the subquestion. Also, D&R Distillation performs the correct reasoning by predicting that `Terry Richardson is older`. These examples highlight the effectiveness of our D&R Distillation method for reasoning interactively with adequately acquired relevant knowledge, which leads to a notably improved performance for knowledge-intensive multi-hop questions.

5 Conclusion

In this paper, we proposed Decompose-and-Response Distillation (D&R Distillation) which enhances the reasoning capabilities of small language models (SLMs) on knowledge-intensive multi-hop question answering. Our approach involves distilling two student models separately, with one student model focusing on decomposing subquestions and another student model focusing on answering subquestions with retrieved background knowledge. Through extensive experiments, we showed that D&R Distillation outperforms previous Chain-of-thought Distillation approaches with much less training data.

Limitations

We conduct experiments on three knowledge-intensive multi-hop question-answering datasets, demonstrating the effectiveness of D&R Distillation. However, our method is specially designed for knowledge-intensive reasoning tasks. This limitation poses a constraint on the wider applicability of our method. We plan to extend D&R Distillation to a wider range of reasoning tasks in the future. On the other hand, due to limitations in computational resources, we were unable to conduct experiments on larger-scale language models (> 1B). We will further explore the performance of D&R Distillation on larger-scale language models in future research.

Ethics Statement

The proposed method has no obvious potential risks. All the scientific artifacts used/created are properly cited/licensed, and the usage is consistent with their intended use. All the data used in this work contains no private information.

Acknowledge

This work was supported by the Strategic Priority Research Program of Chinese Academy of Sciences (No. XDA27020203) and the National Natural Science Foundation of China (No. 62376270, No. 62276264) and OPPO Research Fund.

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.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

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*.

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.

Chengcheng Han, Xiaowei Du, Che Zhang, Yixin Lian, Xiang Li, Ming Gao, and Baoyuan Wang. 2023. DialCoT meets PPO: Decomposing and exploring reasoning paths in smaller language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8055–8068, Singapore. Association for Computational Linguistics.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2023. Large language models are reasoning teachers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14852–14882, Toronto, Canada. Association for Computational Linguistics.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex 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. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017, Toronto, Canada. Association for Computational Linguistics.

Aditya Kalyanpur, Siddharth Patwardhan, BK Boguraev, Adam Lally, and Jennifer Chu-Carroll. 2012. Fact-based question decomposition in deepqa. *IBM Journal of Research and Development*, 56(3.4):13–1.

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*.

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.

Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. 2023a. Symbolic chain-of-thought distillation: Small models can also "think" step-by-step. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2665–2679, Toronto, Canada. Association for Computational Linguistics.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2023b. Making language models better reasoners with step-aware verifier. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5315–5333, Toronto, Canada. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2023. Teaching small language models to reason. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1773–1781, Toronto, Canada. Association for Computational Linguistics.

Pruthvi Patel, Swaroop Mishra, Mihir Parmar, and Chitta Baral. 2022. Is a question decomposition unit all we need? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4553–4569, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online. Association for Computational Linguistics.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711, Singapore. Association for Computational Linguistics.

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.

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, Toronto, Canada. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

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*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037, Toronto, Canada. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

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

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Jianyi Zhang, Aashiq Muhamed, Aditya Anantharaman, Guoyin Wang, Changyou Chen, Kai Zhong, Qingjun Cui, Yi Xu, Belinda Zeng, Trishul Chilimbi, and Yiran Chen. 2023. ReAugKD: Retrieval-augmented knowledge distillation for pre-trained language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1128–1136, Toronto, Canada. Association for Computational Linguistics.

Shen Zheng, Jie Huang, and Kevin Chen-Chuan Chang. 2023. Why does chatgpt fall short in answering questions faithfully? *arXiv preprint arXiv:2304.10513*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

A Implementation Detail

Dataset For HotpotQA and 2WikiMultiHop datasets, we use the official dev split since the test split is not publicly available. For StrategyQA, we split the training set into a 9: 1 ratio to build the in-house test set. Moreover, to control the quality of generated samples, we discard generated D&R Distillation samples if the F1 between the predicted answer and the ground is below 0.7.

Training and Inference For all our experiments, we fine-tune the small language model using the AdamW optimizer (Loshchilov and Hutter, 2019). We fine-tune student models for a maximum of 20 epochs, setting the batch size at 16 and the learning rate at $3e-4$. All our experiments can be run on 2 NVIDIA GTX 3090 GPUs. For text generation, we apply greedy decoding for all models following (Wei et al., 2022; Kojima et al., 2022).

Retriever We use Wikipedia as the external knowledge base and BM25 as the retriever. We set TopK=3 for our retriever, for retrieved passages, we keep the first 100 words for each passage.

B Prompts

Prompting examples for the three datasets can be found on Table 3, Table 4, and Table 5.

Question: What is the elevation range for the area that the eastern sector of
the Colorado orogeny extends into?
Subquestion: What does the eastern sector of the Colorado orogeny extends into?
Intermediate answer: The eastern sector of Colorado orogeny extends into the High Plains.
Subquestion: What is the elevation range for the High Plains?
Intermediate answer: High Plains rise in elevation from around 1,800 to 7,000 ft.
So the final answer is: 1,800 to 7,000 ft

Question: Musician and satirist Allie Goertz wrote a song about the "The Simpsons"
character Milhouse, who Matt Groening named after who?
Subquestion: Who is the "The Simpsons" character Milhouse named after.
Intermediate answer: Richard Milhous Nixon
So the final answer is: Richard Milhous Nixon

Question: Which documentary is about Finnish rock groups, Adam Clayton Powell or The Saimaa Gesture?
Subquestion: What is the documentary Adam Clayton Powell (film) about?
Intermediate answer: Adam Clayton Powell (film) is a documentary about an African-American politician.
Subquestion: What is the documentary The Saimaa Gesture (film) about?
Intermediate answer: The Saimaa Gesture is a film about three Finnish rock groups.
So the final answer is: The Saimaa Gesture

Question: Which magazine was started first Arthur's Magazine or First for Women?
Subquestion: When was Arthurś Magazine started?
Intermediate Answer: Arthurś Magazine was started in 1844.
Subquestion: When was First for Women started?
Intermediate Answer: First for Women was started in 1989.
So the final answer is: Arthurś Magazine

Table 3: Prompts for the HotpotQA dataset.

Question: When did the director of film Hypocrite (Film) die?
Subquestion: Who directed the film Hypocrite (Film)?
Intermediate answer: Miguel Morayta.
Subquestion: When did Miguel Morayta die?
Intermediate answer: Miguel Morayta died on 19 June 2013.
So the final answer is: 19 June 2013

Question: Are both Kurram Garhi and Trojkrsti located in the same country?
Subquestion: Which country is Kurram Garhi located in?
Intermediate answer: Kurram Garhi is located in the country of Pakistan.
Subquestion: Which country is Trojkrsti located in?
Intermediate answer: Trojkrsti is located in the country of Republic of Macedonia.
So the final answer is: No

Question: Which album was released earlier, What'S Inside or Cassandra'S Dream (Album)?
Subquestion: When was the album What's Inside released?
Intermediate answer: What's Inside was released in the year 1995.
Subquestion: When was the album Cassandra'S Dream (Album) released?
Intermediate answer: Cassandra's Dream (album) was released in the year 2008.
So the final answer is: What's Inside

Question: What is the cause of death of Grand Duke Alexei Alexandrovich Of Russia's mother?
Subquestion: Who is the mother of Grand Duke Alexei Alexandrovich of Russia?
Intermediate answer: Maria Alexandrovna.
Subquestion: What is the cause of death of Maria Alexandrovna?
Intermediate answer: Maria Alexandrovna died from tuberculosis.
So the final answer is: Ytuberculosis

Table 4: Prompts for the 2WikiMultiHop dataset.

Question: Could the members of The Police perform lawful arrests?
Subquestion: Who can perform lawful arrests?
Intermediate answer: Only law enforcement officers can perform lawful arrests.
Subquestion: Are members of The Police also?
Intermediate answer: No, The members of The Police were musicians, not law enforcement officers.
So the final answer is: No

Question: Is a Boeing 737 cost covered by Wonder Woman (2017 film) box office receipts?
Subquestion: How much does a Boeing 737 cost?
Intermediate answer: The average cost of a US Boeing 737 plane is 1.6 million dollars.
Subquestion: How much did the 2017 movie Wonder Woman gross?
Intermediate answer: Wonder Woman (2017 film) grossed over 800 million dollars at the box office.
So the final answer is: Yes

Question: Would a Monoamine Oxidase candy bar cheer up a depressed friend?
Subquestion: Depression is caused by low levels of what chemicals?
Intermediate answer: Depression is caused by low levels of serotonin, dopamine and norepinephrine.
Subquestion: Can Monoamine Oxidase lowers levels of serotonin, dopamine and norepinephrine?
Intermediate answer: No, Monoamine Oxidase breaks down neurotransmitters
and lowers levels of serotonin, dopamine and norepinephrine.
So the final answer is: No

Question: Is the language used in Saint Vincent and the Grenadines rooted in English?
Subquestion: What language is used in Saint Vincent and the Grenadines?
Intermediate answer: The primary language spoken in Saint Vincent and the Grenadines is Vincentian Creole.
Subquestion: Is Vincentian Creole based in English?
Intermediate answer: Yes, Vincentian Creole is English-based.
So the final answer is: Yes

Table 5: Prompts for the StrategyQA dataset.