StepER: Step-wise Knowledge Distillation for Enhancing Reasoning Ability in Multi-Step Retrieval-Augmented Language Models

Kyumin Lee[1,2*] Minjin Jeon[1*] Sanghwan Jang[1] Hwanjo Yu[1†]
[1]POSTECH [2]KT
{qmin2, minjinj, s.jang, hwanjoyu}@postech.ac.kr

Abstract

Answering complex real-world questions requires step-by-step retrieval and integration of relevant information to generate well-grounded responses. However, existing knowledge distillation methods overlook the need for different reasoning abilities at different steps, hindering transfer in multi-step retrieval-augmented frameworks. To address this, we propose Step-wise Knowledge Distillation for Enhancing Reasoning Ability in Multi-Step Retrieval-Augmented Language Models (StepER). StepER employs step-wise supervision to align with evolving information and reasoning demands across stages. Additionally, it incorporates difficulty-aware training to progressively optimize learning by prioritizing suitable steps. Our method is adaptable to various multi-step retrieval-augmented language models, including those that use retrieval queries for reasoning paths or decomposed questions. Extensive experiments show that StepER outperforms prior methods on multi-hop QA benchmarks, with an 8B model achieving performance comparable to a 70B teacher model.

1 Introduction

Large language models (LLMs) that employ multi-step retrieval-augmented generation demonstrate strong reasoning abilities for solving complex real-world problems (Trivedi et al., 2022a; Shao et al., 2023; Press et al., 2022; Yao et al., 2023). However, such sophisticated reasoning abilities are primarily observed in large models (Wei et al., 2022; Chung et al., 2024), which incur substantial inference costs. To mitigate this, knowledge distillation (KD) has been introduced to transfer these abilities to smaller models (Hsieh et al., 2023; Mitra et al., 2023; Lee et al., 2024). Most existing KD approaches typically train student models to mimic teacher-generated rationales (Luo et al., 2023; Kang et al., 2023; Yu et al., 2023). Although effective for relatively simple tasks, these methods fall short when handling complex real-world problems.

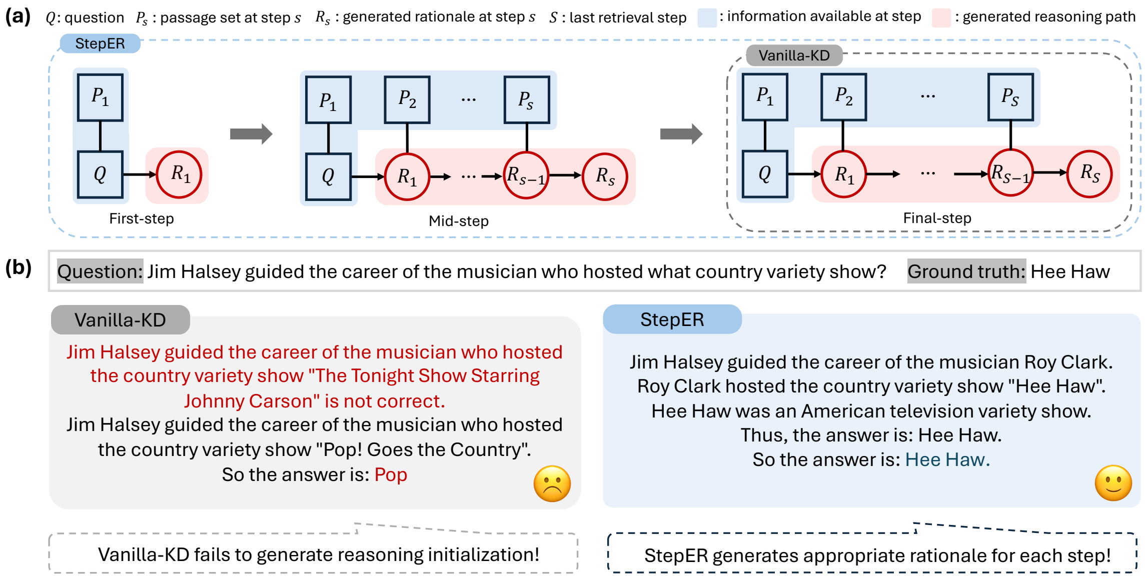
To answer complex questions, a model must develop multiple reasoning abilities. For instance, consider a doctor diagnosing a patient with ankle pain. The diagnostic process can be broken down into three distinct stages: (1) Reasoning Initialization, where the doctor identifies potential conditions based on initial symptoms; (2) Reasoning Expansion, where additional tests, such as X-rays for fractures or ultrasounds for soft tissue damage, are performed to gather more specific information; and (3) Reasoning Aggregation, where the doctor makes a final diagnosis and treatment plan considering all collected information. Similarly, a model needs to learn step-by-step reasoning and adapt to the varying amount of information required at each stage for solving complex problem.

Existing KD methods are limited in these scenarios, as they fail to account for reasoning abilities and the varying amounts of information required at each step (Kang et al., 2023; Luo et al., 2023; Yu et al., 2023; Li et al., 2024). In general, they train the student model to generate the entire reasoning path from the accumulated retrieval results, without considering step-wise differences in information and learning difficulty. As shown in Figure 1, the Vanilla-KD model fails to initialize the reasoning path properly, attempting to generate the entire path in the first-step with minimal information, which limits its performance in multi-step retrieval settings.

To address this limitation, we propose **Step**-wise Knowledge Distillation for **E**nhancing **R**easoning Ability in a Multi-Step Retrieval-Augmented LM (StepER). StepER constructs a step-wise dataset using a teacher multi-step retrieval-augmented LM, enabling the model to learn reasoning abilities specific to each step. First-step data guides the model

[*] Equal contribution.
[†] Corresponding author.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 29501–29523
November 4-9, 2025 ©2025 Association for Computational Linguistics



Question: Jim Halsey guided the career of the musician who hosted what country variety show? Ground truth: Hee Haw

Vanilla-KD:
Jim Halsey guided the career of the musician who hosted the country variety show "The Tonight Show Starring Johnny Carson" is not correct.
Jim Halsey guided the career of the musician who hosted the country variety show "Pop! Goes the Country".
So the answer is: Pop

StepER:
Jim Halsey guided the career of the musician Roy Clark.
Roy Clark hosted the country variety show "Hee Haw".
Hee Haw was an American television variety show.
Thus, the answer is: Hee Haw.
So the answer is: Hee Haw.

Vanilla-KD fails to generate reasoning initialization!

StepER generates appropriate rationale for each step!

Figure 1: Comparison of Vanilla-KD and STEPER. (a) illustrates the conceptual differences in training data. Unlike Vanilla-KD which only uses final-step data, STEPER leverages data from all reasoning stages—first-step (initial reasoning based on the first retrieved passages), mid-step (intermediate reasoning with accumulated information), and final-step (complete reasoning with all retrieved passages). STEPER learns reasoning abilities more effectively by leveraging all steps of reasoning data during training. (b) presents answer examples from both models. Vanilla-KD often fails in early reasoning stages and generates incorrect answers, whereas STEPER performs coherent reasoning throughout and reaches the correct answer.

to initiate reasoning from limited initial information, mid-step data helps expand reasoning by incorporating additional retrieved evidence, and final-step data enables the model to aggregate and conclude based on the complete context. This approach allows the model to acquire reasoning capabilities for complex questions while considering the information required at each step.

We also introduce reasoning difficulty-aware training to further enhance the model's reasoning abilities. The model initially focuses on tasks that are easier to learn, gradually shifting toward more challenging ones as training progresses. This adaptive strategy allows the model to optimize learning based on its current capabilities, resulting in improved reasoning performance. As shown in Figure 1, a model trained with STEPER successfully identifies the artist, the show hosted by the artist, the country where it aired, and ultimately produces the correct answer.

STEPER offers several advantages for answering complex questions. First, it outperforms Vanilla-KD methods, with experiments showing an average accuracy improvement of approximately 9.5%. G-Eval results confirm that step-wise reasoning is crucial for enhancing reasoning abilities. Second, STEPER is flexible and can be applied to various multi-step retrieval-augmented LM frameworks. Further, STEPER is model-scalable, achieving performance comparable to a 70B teacher model with a 8B model.

Our main contributions are as follows: (1) We categorize the essential reasoning abilities required for multi-retrieval settings and demonstrate the need for methods to enhance each ability. (2) We propose STEPER, a method that uses step-wise data and reasoning difficulty-aware training to effectively learn the necessary reasoning abilities. (3) Extensive analyses show that STEPER outperforms existing KD approaches, improving both overall performance and scalability across various model sizes.

2 Related Work

Retrieval-Augmented LM Retrieval-augmented LMs have significantly improved performance in knowledge-intensive tasks such as Open-Domain Question Answering (Lewis et al., 2020; Guu et al., 2020). These models typically consist of a retriever that selects relevant documents and a generator that constructs responses based on the retrieved information (Borgeaud et al., 2022; Izacard et al., 2023; Shi et al., 2023). To answer based on documents most relevant to the question, Kim et al. (2024), Xu et al. (2023) have explored approaches that refine retrieved documents before generation, by summarizing evidence. However, Jiang et al. (2024) shows that improving the quality of retrieval results alone remains insufficient for multi-hop QA tasks, indicating the need for more effective methods to facilitate complex reasoning in question answering.

Multi-Step Retrieval-Augmented LM To address the limitations of single-step retrieval in handling complex queries, multi-step retrieval-augmented LMs have been introduced (Trivedi et al., 2022a; Shao et al., 2023; Jeong et al., 2024). These models iteratively retrieve information throughout the reasoning process. Trivedi et al. (2022a), Shao et al. (2023) leverage previously generated rationales as queries for subsequent retrieval, while Press et al. (2022) decomposes the original question into sub-questions and answers them independently.

KD for Retrieval-Augmented LM Several studies have explored the use of teacher-generated rationales to improve the training of retrieval-augmented LMs (Xu et al., 2024). In addition to simply utilizing teacher rationales, recent studies have been proposed to enhance search result quality using rationales (Kang et al., 2023) or to improve answer generation by reflecting the relevance between the retrieved passages and the question (Luo et al., 2023; Yu et al., 2023). However, these methods primarily focus on single-step retrieval settings, which limits their performance in multi-hop question answering tasks.

Recently, Asai et al. (2023) has been introduced to enhance the training of multi-step retrieval-augmented LMs by learning when to retrieve and which documents to incorporate into responses. This approach focuses on integrating high-quality search results into answers but overlooks the step-wise reasoning abilities needed for complex questions and requires additional models for training, increasing the cost.

3 Preliminaries

We formalize retrieval-augmented generation (RAG) in the context of multi-step reasoning. Specifically, let q denote the original input question, and let the reasoning process proceed over S steps. During the first step, the model generates an initial reasoning r_1. In the following $S-2$ steps, it expands its reasoning through intermediate outputs $\{r_2, r_3, \ldots, r_{S-1}\}$. Finally, in the S-th step, it produces the answer, denoted by, $r_S = a$.

Single-Step RAG In the single-step RAG, the model accesses an external knowledge source only once before generating both its reasoning chain and final answer. Let P_1 be the top-K passages retrieved from the knowledge source given the original question q. The generation process is then factorized as

$$P\big(R \mid q, P_1\big) \,\cdot\, P\big(a \mid q, P_1, R\big). \tag{1}$$

Here, the model first generates the intermediate reasoning steps R conditioned on $\{q, P_1\}$, and then produces the final answer a based on $\{q, P_1, R\}$. Although this approach simplifies the pipeline, previous works have demonstrated that it is inadequate for complex multi-hop queries that require additional (Trivedi et al., 2022a; Jeong et al., 2024; Gao et al., 2023; Shao et al., 2023; Jiang et al., 2023).

Multi-Step RAG Multi-step RAG extends single-step RAG by iteratively retrieving new passages over multiple steps. At step s, let q_s be a step-search query, which is constructed based on the partial chain of reasoning $R_{<s} = \{r_1, \ldots, r_{s-1}\}$. Using q_s to query the external knowledge source, we retrieve the top-K relevant passages P_s. We denote by $P_{\leq s} = \bigcup_{i=1}^{s} P_i$ the collection of all passages retrieved up to step s. For S total steps, the generation process is factorized as

$$\left[\prod_{s=1}^{S-1} P\big(r_s \mid q,\, P_{\leq s},\, R_{<s}\big)\right] \cdot P\big(a \mid q,\, P_{\leq S},\, R_{<S}\big), \tag{2}$$

By repeatedly retrieving and integrating new evidence, Multi-step RAG is naturally suited to address complex or multi-hop questions.

4 StepER Framework

We propose a novel framework, **StepER**, to enhance the step-specific reasoning abilities of student models. Our framework comprises two main stages: a data construction phase, where a teacher model generates a step-wise dataset, and a training phase, where a student model is trained on this data using a reasoning difficulty-aware learning method.

4.1 Data Construction

Based on Equation 2, we propose that the accessible information in a multi-step RAG increases with each step, creating distinct reasoning demands. To

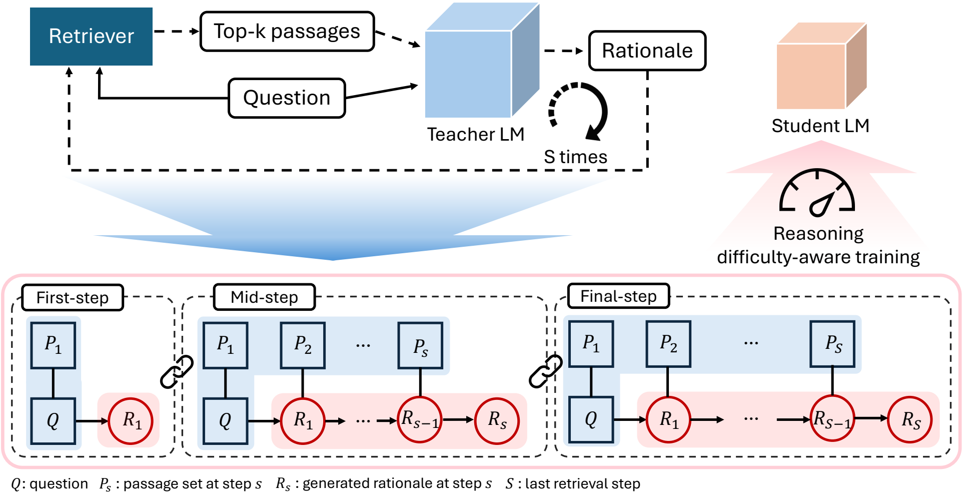

Retriever
Top-k passages
Question
Teacher LM
S times
Rationale
Student LM
Reasoning difficulty-aware training
First-step
Mid-step
Final-step
P_1
P_2
P_s
Q
R_1
R_{s-1}
R_s
Q: question P_s : passage set at step s R_s : generated rationale at step s S : last retrieval step

Figure 2: Overview of the STEPER framework. We use a teacher LM to construct the dataset via multi-step retrieval, and train the student model with a difficulty-aware strategy that prioritizes reasoning steps more suitable for learning.

reflect this, we divide the reasoning process into three stages: *initialization*, *expansion*, and *aggregation*, which align with human problem-solving theory (Simon and Newell, 1971).

Specifically, the initialization stage involves reasoning with minimal information to establish a starting point. The expansion stage focuses on identifying and retrieving additional information based on prior reasoning, while the aggregation stage integrates all collected evidence to produce a final answer. These stages require distinct reasoning abilities.

The student learns these three reasoning skills from a teacher, via a step-wise dataset, denoted as D_{steps}, constructed from the original dataset D. Given a complex QA dataset $\mathcal{D} = \{(q^{(i)}, a^{(i)})\}_{i=1}^{n}$, where each $q^{(i)}$ is a question and $a^{(i)}$ is its correct answer, we construct a stepwise dataset $\mathcal{D}_{\text{steps}}$ in which every sample explicitly records multiple intermediate reasoning steps with the corresponding accessible information.

Reasoning Initialization For each question $q^{(i)}$, we retrieve first passages $P_1^{(i)}$ by querying an external knowledge source with $q^{(i)}$. We then prompt the teacher model $\mathcal{T}$ to produce the initial reasoning step $r_1^{(i)}$ from $\left(q^{(i)}, P_1^{(i)}\right)$. We retain the initial reasoning step r_1 and then proceed to the next step.

Reasoning Expansion Based upon the initial rationale, we prompt the teacher model $\mathcal{T}$ to generate the next reasoning step. Specifically, at step $s > 1$, we retrieve additional passages $P_s^{(i)}$ using q_s as a step-search query. A step-search query is derived from a partial reasoning chain of the teacher $R_{\leq s-1} = \{r_1, \ldots, r_{s-1}\}$. This can be the form of a previous reasoning step or decomposed question. Then, the cumulative information $\left(q^{(i)}, P_{\leq s}^{(i)}, R_{\leq s-1}^{(i)}\right)$ is provided as input, from which $\mathcal{T}$ produces the next reasoning step $r_s^{(i)}$. This iterative process continues up to a maximum of $S - 1$ steps. If at any point $r_s^{(i)}$ includes the answer flag (e.g., beginning with "`So the answer is:`"), we record the reasoning chain constructed up to the previous step and terminate the expansion step early.

Reasoning Aggregation Upon reaching the last step S or an early termination in the expansion step, we prompt $\mathcal{T}$ to aggregate all previous reasoning steps and passages. Concretely, $\mathcal{T}$ is instructed to append a concluding statement like "`So the answer is:`" and explicitly provide $a^{(i)}$.

Filtering Dataset After generating all reasoning steps for each $(q^{(i)}, a^{(i)})$, we filter out samples where the teacher's final statement does not match the ground truth $a^{(i)}$, ensuring that $\mathcal{D}_{\text{steps}}$ only contains the correct reasoning processes. Ultimately, every sample in $\mathcal{D}_{\text{steps}}$ illustrates how $\mathcal{T}$ **(i)** *initializes* reasoning from limited context, **(ii)** *expands* partial reasoning with newly retrieved evidence, and **(iii)** *aggregates* all partial results into a final answer, corresponding respectively to the First-step, Mid-step, and Final-step categories illustrated in Figure 2.

4.2 Learning Objectives

Multi-task Learning We train the student model $\mathcal{M}$ on the stepwise dataset $\mathcal{D}_{\text{steps}}$ to distill multi-step reasoning abilities. Formally, we minimize the following objective:

$$\mathcal{L} = \frac{1}{3n}\sum_{i=1}^{n}\Big[\underbrace{\ell\big(\mathcal{M}(q^{(i)}, P^{(i)}_{\le 1}),\, R^{(i)}_{\le 1}\big)}_{\text{(a) reasoning initialization}} + \underbrace{\sum_{s=2}^{S-1}\ell\big(\mathcal{M}(q^{(i)}, P^{(i)}_{\le s}),\, R^{(i)}_{\le s}\big)}_{\text{(b) reasoning expansion}} + \underbrace{\ell\big(\mathcal{M}(q^{(i)}, P^{(i)}_{\le S}),\, (R^{(i)}_{<S}\|a^{(i)})\big)}_{\text{(c) reasoning aggregation}}\Big], \tag{3}$$

where $\ell(\cdot, \cdot)$ is the cross-entropy loss between predicted and target tokens, n is the total number of samples, and $\|$ in (c) denotes string concatenation.

Reasoning Difficulty-Aware Training As training progresses, the model's perception of step difficulty evolves, requiring a learning strategy that continuously adapts to its changing capabilities. (Liang and Zhang, 2020; Guo et al., 2018; Murugesan and Carbonell, 2017). To this end, we employ an adaptive weighting scheme (Kendall et al., 2017; Chen et al., 2021) that dynamically adjusts training priorities, allowing the model to focus on the most appropriate steps at each stage. We represent the difficulty of each reasoning task as a trainable parameter σ. In Equation (3), (a), (b), and (c) correspond to L_{init}, L_{exp}, and L_{agg} respectively; then, the final objective is formulated as:

$$\mathcal{L}_{\text{final}} = \sum_{j \in \{\text{init, exp, agg}\}} \left(\frac{1}{2\,\sigma_j^2}\, L_j \;+\; \log\, \sigma_j\right), \tag{4}$$

where $\log \sigma_j$ acts as a regularization term. During training, more challenging tasks are guided to have higher σ values, while easier tasks have lower ones. This allows the model to dynamically reweight its learning focus based on the perceived difficulty of each step, ultimately facilitating more effective multi-step reasoning.

5 Experiments

5.1 Experimental Setup

Backbone Model We use Llama3.1-Instruct 70B (Dubey et al., 2024) as our teacher model $\mathcal{T}$, with Llama3.1-Instruct 8B as the student model $\mathcal{M}$. Unless otherwise specified, all baseline methods employ Llama3.1-Instruct.

Datasets and Metrics We evaluate on three widely used multi-hop QA benchmarks that involve complex queries: 2WikiMultiHopQA (2Wiki) (Ho et al., 2020), HotpotQA (Yang et al., 2018), and MuSiQue (Trivedi et al., 2022b) that are recognized for requiring more complex and multi-step reasoning (Welbl et al., 2018; Yang et al., 2018). We report Exact Match (EM), F1, and Accuracy (Acc), where Acc measures whether the ground-truth answer is present in the model's generated text.

Baselines We compare a wide range of retrieval-augmented generation (RAG) methods that cover both few-shot in-context learning (ICL) and knowledge distillation, while varying the number of retrieval times (*No*, *Single*, or *Multiple*).

In ICL, we include a non-retrieval few-shot baseline for reference, since LLMs already encode a large amount of knowledge (Zhao et al., 2023). Next, we evaluate Vanilla-RAG (Lewis et al., 2020) and SuRE (Kim et al., 2024) under the single-step retrieval setting. Vanilla-RAG retrieves relevant documents and generates an answer conditioned on the retrieved context. For multi-step retrieval in ICL, we compare two ways of updating the step-search query: one in which the query is updated with previously generated context, as in ITER-RETGEN (Shao et al., 2023) and IRCOT (Trivedi et al., 2022a), and another where the model decomposes the original question into multiple sub-queries, as in Self-Ask (Press et al., 2022) and ReAct (Yao et al., 2023).

In knowledge distillation, we compare StepER with several baselines, including SAIL (Luo et al., 2023), KARD (Kang et al., 2023), CoN (Yu et al., 2023), Self-RAG (Asai et al., 2023), and Vanilla-KD. Vanilla-KD trains on the complete multi-step reasoning process as shown in Figure 1-(a). Given the question and all retrieved passages, it is supervised to generate the entire reasoning path and final answer. In contrast, our method provides stepwise supervision conditioned only on the passages retrieved so far. Further details on additional ICL and knowledge distillation baselines are provided in Appendix A.2.

Implementation Details We follow the corpus selection and data preprocessing setup from the previous work Trivedi et al. (2022a). For passage retrieval, we adopt an off-the-shelf retriever BM25 (Robertson and Zaragoza, 2009) with a maximum of $S = 5$ retrieval steps, retrieving the top-$K = 4$

Retrieval steps		2Wiki			HotpotQA			MuSiQue			Avg.		
		EM	F1	Acc	EM	F1	Acc	EM	F1	Acc	EM	F1	Acc
						In-Context Learning							
No	Llama3.1 8B	29.83	35.59	33.69	29.18	38.76	35.01	8.68	17.91	13.22	22.56	30.75	27.31
	Llama3.1 70B	45.47	51.09	47.89	40.61	51.25	45.86	16.19	25.94	23.28	34.09	42.76	39.01
	GPT-4o-mini	25.51	40.80	27.09	28.15	41.25	35.81	11.84	24.03	15.89	21.83	35.36	26.26
	GPT-4o	52.26	65.88	53.70	40.69	57.24	48.05	21.62	35.22	28.50	38.19	52.78	43.42
Single	Vanilla-RAG 8B	35.97	43.10	38.88	38.25	49.08	46.15	11.18	20.91	22.57	28.46	37.69	35.86
	Vanilla-RAG 70B	51.01	57.80	53.83	45.25	56.30	52.93	19.84	30.79	31.58	38.70	48.29	46.08
	SuRE 70B	25.20	41.34	41.20	30.60	48.23	41.00	11.60	22.00	19.40	22.46	37.19	33.86
Multi	ITER-RETGEN3 70B	44.60	50.92	46.20	48.20	60.12	53.40	24.20	33.17	30.00	39.00	48.07	43.20
	ITER-RETGEN4 70B	44.20	50.54	45.60	49.40	60.92	54.60	24.80	32.98	30.40	39.46	48.14	43.53
	ITER-RETGEN5 70B	44.00	50.35	45.60	49.40	60.51	54.80	24.00	31.92	29.60	39.13	47.59	43.33
	IRCOT 8B	41.80	49.94	44.80	43.40	53.82	50.80	17.20	27.57	28.40	34.13	43.77	41.33
	IRCOT 70B	60.16	67.06	62.37	49.60	61.31	57.23	24.30	35.29	34.74	44.68	54.55	51.45
	Self-Ask 8B	38.80	47.41	43.00	40.80	52.00	48.20	15.83	23.58	23.85	31.81	41.00	38.35
	Self-Ask 70B	57.80	66.44	61.00	50.60	62.60	59.40	25.20	36.68	33.80	44.53	55.24	51.40
	ReAct 8B	40.20	49.50	43.00	33.60	43.96	39.60	14.80	24.73	21.20	29.53	39.40	34.60
	ReAct 70B	59.40	68.58	61.60	46.00	59.89	53.40	**28.20**	**39.46**	**35.60**	44.53	55.98	50.20
						Knowledge Distillation							
Single	SAIL	47.90	54.06	49.50	44.56	56.30	51.41	6.41	16.34	10.62	32.96	42.23	37.18
	KARD	47.80	54.48	51.40	43.80	54.59	53.00	14.60	25.54	24.60	35.40	44.87	43.00
	CoN	45.66	53.93	48.89	42.46	53.34	51.00	16.36	26.85	25.86	34.96	44.70	41.91
Multi	Self-RAG	41.15	46.99	42.82	36.85	44.88	41.26	9.16	17.19	12.80	29.05	36.35	32.29
	Vanilla-KD	60.06	65.55	62.16	46.40	57.28	54.80	20.92	32.46	30.13	42.46	51.76	49.03
	StepER	**63.60**	**69.45**	**66.00**	**51.00**	**62.80**	**61.00**	23.59	36.13	34.07	**46.06**	**56.12**	**53.69**

Table 1: Overall experimental results with **Llama3.1-Instruct** as the base model. The table is categorized by retrieval steps *No*, *Single*, and *Multi*, which indicate how many times retrieval is performed during the generation of the full reasoning path. All listed models (SAIL, KARD, CoN, Self-RAG, and Vanilla-KD) are trained with Llama3.1-Instruct 8B under the Knowledge Distillation criteria. Averages (Avg.) are computed across three datasets: 2Wiki, HotpotQA, and MuSiQue. The number for ITER-RETGEN represents the maximum number of retrieval steps.

passages at each step. We train the models using a learning rate of 5×10^{-6} for total 2 epochs, along with a cosine scheduler and linear warmup. Experiments run on 4×A100 GPUs with DeepSpeed ZeRO Stage 3 and gradient checkpointing to reduce memory consumption.

5.2 Main Results

Table 1 shows the performance of various approaches on 2Wiki, HotpotQA, and MuSiQue with Llama3.1-Instruct. We first note that single-time retrieval methods struggle to address complex queries, and even recent improvements (Kim et al., 2024) exhibit a noticeable gap compared to multi-time retrieval. In addition, an accuracy gap persists between 8B and 70B models under multi-step RAG ICL, highlighting the importance of model size in complex reasoning tasks.

StepER stands out as it delivers the best performance among knowledge distillation methods, achieving a 9.5% average accuracy improvement over Vanilla-KD and outperforming all baselines on 2Wiki and HotpotQA. These results underscore how StepER effectively inherits step-wise reasoning abilities from the teacher model, enabling strong reasoning performance with a smaller student model.

6 Analysis

6.1 Effectiveness of Step Data in Enhancing Reasoning Abilities

We conduct an experiment to evaluate the effectiveness of step data in enhancing reasoning abilities required for multi-step retrieval-augmented LM. We categorize the necessary reasoning abilities into three types for evaluation: (1) *Reasoning Initialization*, (2) *Reasoning Expansion*, and (3) *Reasoning Aggregation*, as described in Section 4. To evaluate these abilities, we perform binary classification for each criterion using GPT-4o, evaluated on the HotpotQA dataset. The detailed prompt used for evaluation is provided in the Appendix D. We train the models using various step data configurations, specifically: Vanilla-KD (S=5), Vanilla-KD+First-step (S=1,5), Vanilla-KD+First-step+First Mid-step (S=1,2,5), and StepER (all step data), with a maximum of $S = 5$ retrieval steps.

Vanilla-KD relies solely on Final-step data and struggles to capture detailed intermediate reasoning. In contrast, adding First-step data strengthens the ability to initiate reasoning (*Reasoning Initialization*). By offering a clear starting point for multi-step reasoning, the model can more effectively identify and focus on relevant information at the beginning of the reasoning process. Furthermore, incorporating the First-step data and First Mid-step data improves the expansion process (*Reasoning Expansion*), enabling the model to elaborate on its initial line of reasoning before arriving at the final conclusion. Finally, STEPER, which jointly leverages all step data, outperforms all other models, confirming that step-wise data enhances the reasoning abilities required for multi-step retrieval settings.

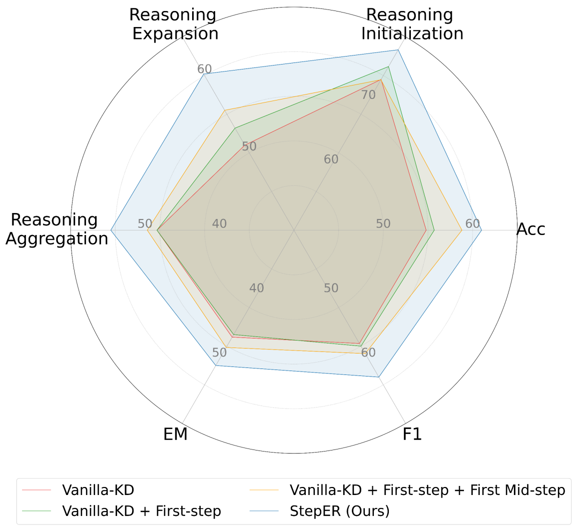


Figure 3: GPT evaluation results on HotpotQA across three reasoning stages under different step data configurations. STEPER, which utilizes all available step data, achieves the highest performance across all evaluation criteria, demonstrating the effectiveness of step-wise training for multi-step retrieval.

6.2 Effectiveness of Difficulty-Aware Adaptive Weighting Strategy

Strategy	HotpotQA			MuSiQue		
	EM	F1	Acc	EM	F1	Acc
Uniform ($\lambda = 1, 1, 1$)	50.40	61.57	58.40	21.67	33.28	33.58
Weight First ($\lambda = 1.5, 1, 0.5$)	49.10	61.63	57.70	21.04	31.24	32.46
Weight Last ($\lambda = 0.5, 1, 1.5$)	48.80	60.78	58.00	21.91	33.85	33.37
Difficulty-Aware (Ours)	**51.00**	**62.80**	**61.00**	**23.59**	**36.13**	**34.07**

Table 2: Comparison of our Difficulty-Aware adaptive weighting strategy against several fixed-weight baselines. Our Difficulty-aware approach achieves consistently higher performance by dynamically adjusting training focus according to the relative difficulty of each reasoning step.

As introduced in Equation (4), our overall loss consists of three partial losses $\{L_{\text{init}}, L_{\text{exp}}, L_{\text{agg}}\}$, each scaled by $\frac{1}{2\sigma_j^2}$. Specifically, We set $\lambda_j = \frac{1}{2\sigma_j^2}$. adaptively control the relative difficulty of each task. Table 2 compares this Difficulty-Aware Adaptive strategy against several fixed-weight baselines. Notably, we observe consistent improvements on both HotpotQA and MuSiQue. These results demonstrate that the model benefits from dynamically allocating attention to the most learnable step at each training phase. The relative change of σ_j over training is illustrated in Appendix C, Figure 7.

6.3 Applicability to Another Multi-step Retrieval Approach

We further investigate the generality of our step-wise knowledge distillation by integrating STEPER with Self-Ask, another multi-step retrieval framework where each step-search query is generated from a decomposition of the original question. As shown in Table 3, STEPER consistently outperforms Vanilla-KD on both HotpotQA and MuSiQue, highlighting the effectiveness of explicitly distilling intermediate rationales at each retrieval step rather than depending only on final-step supervision. In addition, STEPER achieves substantial improvements over the Self-Ask 8B baseline, boosting its accuracy by 9.6% on HotpotQA and 14.95% on MuSiQue. Consequently, these results demonstrate that our approach can integrate seamlessly with various multi-step retrieval-augmented LMs.

Model	HotpotQA			MuSiQue		
	EM	F1	Acc	EM	F1	Acc
Self-Ask 8B	40.80	52.50	48.20	15.83	23.58	23.85
Vanilla-KD	47.60	60.11	56.40	26.90	38.92	37.00
STEPER	49.80	62.33	57.80	**28.20**	**40.52**	**38.80**
Self-Ask 70B	**50.60**	**62.60**	**59.40**	25.20	36.68	33.80

Table 3: Evaluation results of Self-Ask on HotpotQA and MusiQue. We compare the teacher model (Self-Ask 70B) with student models (8B) distilled through either Vanilla-KD or STEPER.

6.4 Model Scalability

Figure 4 shows that STEPER consistently achieves the highest accuracy across all Qwen2.5-Instruct model sizes (0.5B, 1.5B, 3B, and 7B) (Yang et al., 2024) on HotpotQA. Notably, the STEPER 3B model nearly matches the performance of the Qwen2.5-Instruct 72B teacher, while the STEPER

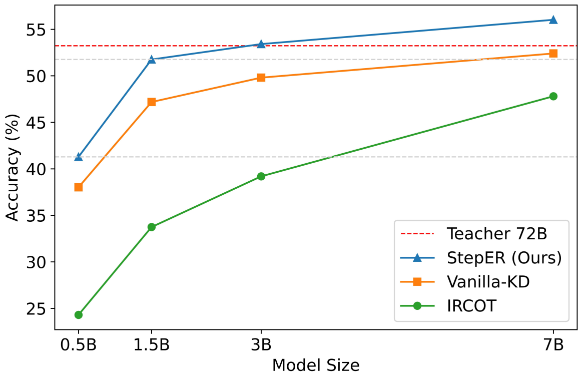


Figure 4: Model scalability of StepER on HotpotQA using Qwen2.5-Instruct. We compare models of varying sizes and demonstrate that StepER scales effectively with consistently strong multi-step reasoning performance.

7B even surpasses it. Furthermore, StepER 3B outperforms the Vanilla-KD 7B, and StepER 1.5B surpasses the Vanilla-KD 3B, indicating that StepER can effectively bridge model-scale gaps by distilling step-wise reasoning abilities. These findings underscore the practicality of StepER in resource-constrained scenarios, where smaller models can achieve performance levels comparable to much larger counterparts (Sanh, 2019; Liu et al., 2024).

6.5 Rationale Validity

To further evaluate the validity of rationales generated by StepER, we use the SubQA dataset (Tang et al., 2021), which consists of original questions paired with corresponding sub-questions. Sub-questions are designed to probe whether the rationale contains sufficient information to answer the original question. The quality of rationales is evaluated by prompting GPT-4o. The detailed prompt used for evaluation is provided in the Appendix D.

We introduce two evaluation criteria: *Sub-question Answerability (SQA)* and *Reasoning Integrity (RI)*. For *SQA*, we assign a score of 1 if both sub-questions can be answered from the rationale, 0.5 if only one can be answered, and 0 otherwise. The average scores are reported as percentages. *RI* measures the percentage of instances where all sub-questions are answerable given that the original question is answered correctly. In addition, we report the accuracy for the original question in SubQA as a reference.

StepER achieves the highest performance in both *SQA* and *RI*, indicating that it generates more valid and coherent rationales for complex questions. Notably, its step-wise rationales consistently include sufficient information to answer corresponding sub-questions. These results demonstrate the effectiveness of step-wise supervision in producing valid intermediate reasoning that supports the final answer.

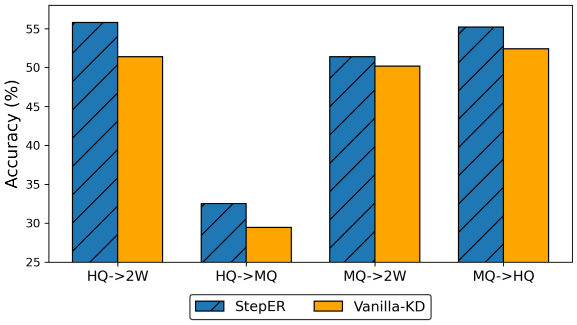


Figure 5: Out-of-domain adaptation results for StepER versus Vanilla-KD across four domain transfer scenarios: HQ→2W, HQ→MQ, MQ→2W, and MQ→HQ. StepER consistently outperforms Vanilla-KD, demonstrating stronger cross-domain generalization.

Model	**Accuracy**	**SQA**	**RI**
IRCOT 70B	**56.76**	<u>70.35</u>	<u>86.00</u>
IRCOT 8B	48.35	66.70	<u>86.00</u>
Vanilla-KD	50.30	62.95	81.00
StepER	<u>55.70</u>	**72.00**	**87.90**

Table 4: Comparison of rationale validity from different models. StepER achieves the highest scores in SQA and RI, highlighting the effectiveness of step-wise supervision in generating valid and coherent rationales across each reasoning steps.

6.6 Out-of-Domain Adaptation

To evaluate the transferability of our approach, we conducted out-of-domain experiments by training the model on one dataset and testing it on another. We use the 2Wiki (2W), HotpotQA (HQ), and MuSiQue (MQ) datasets. Figure 5 shows the performance accuracy of the two methods, StepER and Vanilla-KD, across four domain adaptation scenarios: HQ→2W, HQ→MQ, MQ→2W, and MQ→HQ.

StepER consistently outperforms Vanilla-KD across all four domain adaptation scenarios. In each case, StepER exhibits higher accuracy than Vanilla-KD, with gains ranging from 1% to 4%. This indicates that distilling the teacher's step-wise reasoning ability offers better transferability to out-of-domain tasks than relying solely on the aggregation step as in Vanilla-KD. Overall, these results suggest that learning step-specific reasoning abilities not only enhances in-domain reasoning but also leads to stronger generalization to unseen datasets.

Question: In what european city is a location of the college from which the woman known as Comic Book Girl 19 received her degree? **Answer:** Lacoste, France

Retrieved document #1:

Wikipedia Title: Comic Book Girl 19

Danika Lee Massey, also known as Comic Book Girl 19 or CBG19, is a YouTube personality known for her commentaries on comics, films, books, and television shows. She has a degree in sequential art from the Savannah College of Art and Design.

Retrieved document #2:

Wikipedia Title: Savannah College of Art and Design

Savannah College of Art and Design (SCAD), is a private, nonprofit, accredited university with locations in Savannah, Georgia; Atlanta, Georgia; Hong Kong; and Lacoste, France.

(Other retrieved documents are omitted.)

Vanilla-KD: The woman known as Comic Book Girl 19 received her degree from the Savannah College of Art and Design. The Savannah College of Art and Design is located in Savannah, Georgia. Savannah, Georgia is in the United States. So the answer is: Savannah, Georgia

StepER(Ours): The woman known as Comic Book Girl 19 received her degree from the Savannah College of Art and Design. The Savannah College of Art and Design is located in Savannah, Georgia, United States. However, the Savannah College of Art and Design also has locations in Hong Kong and Lacoste, France. So the answer is: Lacoste, France

Table 5: Qualitative Analysis. In comparison to Vanilla-KD, our StepER demonstrates the ability to expand reasoning and extract more relevant information from the question, resulting in a more accurate answer, as shown in the HotpotQA example. Blue indicates correctly retrieved or referenced information, while red indicates incorrect or misleading references.

6.7 Qualitative Analysis

Table 5 presents a HotpotQA example that highlights the difference between Vanilla-KD and StepER. Vanilla-KD answers correctly when asked about the identity of 'Comic Book Girl 19' and the university she graduated from, but fails to incorporate the step-specific reasoning needed to identify the university's European location, instead returning an incorrect country. In contrast, StepER successfully identifies all relevant details, from 'Comic Book Girl 19' and her university to its European location, producing the correct final answer.

This stepwise behavior is crucial: the initialization step recalls basic facts (e.g., "Comic Book Girl 19's degree location"), the expansion step narrows down to Savannah College's location, and later steps pinpoint the European site. Each step has distinct informational constraints, and our decomposed loss effectively enforces these step-specific reasoning behaviors, enabling StepER to outperform Vanilla-KD in multi-retrieval settings.

7 Conclusion

We propose StepER, a framework designed to enhance the reasoning capabilities of multi-step retrieval-augmented language models. StepER explicitly decomposes the reasoning process into three stages as initialization, expansion, and aggregation, taking into account the distinct characteristic and the information available at each stage. Furthermore, it incorporates a difficulty-aware learning strategy that dynamically adjusts training focus according to the relative complexity of each stage, ensuring effective distillation of reasoning abilities.

Extensive experiments across various model sizes and multi-step retrieval settings demonstrate that StepER consistently improves both overall reasoning performance and the quality of generated reasoning paths. Importantly, it is broadly compatible with a wide range of retrieval-augmented frameworks and scales with different model sizes. These results suggest that StepER provides a promising solution for training smaller models to tackle complex, real-world reasoning tasks, thereby bridging the gap between model efficiency and advanced reasoning capabilities.

Limitations

While StepER effectively enhances the reasoning abilities of multi-step retrieval-augmented LMs, limitations inherent to knowledge distillation still exist. Since the student model learns from the teacher model's rationale, it is crucial to filter the training data to prevent propagating errors. In this study, we filter examples based solely on the correctness of the final answer. However, this method does not penalize wrong reasoning paths that coincidentally lead to a correct answer. We suggest that future work could explore more fine-grained, stepwise filtering based on the validity of the reasoning path, ensuring that the student model trains valid and robust reasoning process rather than relying on shortcuts. Furthermore, leveraging parameter-efficient fine-tuning methods with our proposed method could improve training efficiency, making the framework more practical.

Ethical Considerations

We used publicly available datasets, including 2WikiMultiHotpotQA, HotpotQA, and MuSiQue. For models, we employed publicly released LLaMA-3.1-Instruct, Qwen-2.5-Instruct, GPT-4o, and GPT-4o-mini. Therefore, we do not anticipate significant ethical concerns from our work.

Acknowledgement

This work was supported by Samsung Research Funding & Incubation Center of Samsung Electronics under Project Number SRFC-IT2402-05.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Shijie Chen, Yu Zhang, and Qiang Yang. 2021. Multi-task learning in natural language processing: An overview. *ACM Computing Surveys*.

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

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

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Michelle Guo, Albert Haque, De-An Huang, Serena Yeung, and Li Fei-Fei. 2018. Dynamic task prioritization for multitask learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

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, 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*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C Park. 2024. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. *arXiv preprint arXiv:2403.14403*.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.

Zhouyu Jiang, Mengshu Sun, Lei Liang, and Zhiqiang Zhang. 2024. Retrieve, summarize, plan: Advancing multi-hop question answering with an iterative approach. *arXiv preprint arXiv:2407.13101*.

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. *Advances in Neural Information Processing Systems*, 36:48573–48602.

Alex Kendall, Yarin Gal, and Roberto Cipolla. 2017. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7482–7491.

Jaehyung Kim, Jaehyun Nam, Sangwoo Mo, Jongjin Park, Sang-Woo Lee, Minjoon Seo, Jung-Woo Ha, and Jinwoo Shin. 2024. Sure: Summarizing retrievals using answer candidates for open-domain qa of llms. *arXiv preprint arXiv:2404.13081*.

Hojae Lee, Junho Kim, and SangKeun Lee. 2024. Mentor-kd: Making small language models better multi-step reasoners. *arXiv preprint arXiv:2410.09037*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Xiang Li, Shizhu He, Fangyu Lei, JunYang JunYang, Tianhuang Su, Kang Liu, and Jun Zhao. 2024. Teaching small language models to reason for knowledge-intensive multi-hop question answering. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7804–7816.

Sicong Liang and Yu Zhang. 2020. A simple general approach to balance task difficulty in multi-task learning. *Preprint*, arXiv:2002.04792.

Zechun Liu, Changsheng Zhao, Forrest Iandola, Chen Lai, Yuandong Tian, Igor Fedorov, Yunyang Xiong, Ernie Chang, Yangyang Shi, Raghuraman Krishnamoorthi, et al. 2024. Mobilellm: Optimizing sub-billion parameter language models for on-device use cases. *arXiv preprint arXiv:2402.14905*.

Hongyin Luo, Yung-Sung Chuang, Yuan Gong, Tianhua Zhang, Yoon Kim, Xixin Wu, Danny Fox, Helen Meng, and James Glass. 2023. Sail: Search-augmented instruction learning. *arXiv preprint arXiv:2305.15225*.

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, et al. 2023. Orca 2: Teaching small language models how to reason. *arXiv preprint arXiv:2311.11045*.

Keerthiram Murugesan and Jaime G. Carbonell. 2017. Self-paced multitask learning with shared knowledge. In *International Joint Conference on Artificial Intelligence*.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2022. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*.

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

V Sanh. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. *arXiv preprint arXiv:2305.15294*.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Herbert A Simon and Allen Newell. 1971. Human problem solving: The state of the theory in 1970. *American psychologist*, 26(2):145.

Yixuan Tang, Hwee Tou Ng, and Anthony Tung. 2021. Do multi-hop question answering systems know how to answer the single-hop sub-questions? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3244–3249, Online. Association for Computational Linguistics.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022a. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *arXiv preprint arXiv:2212.10509*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022b. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

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

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2023. Recomp: Improving retrieval-augmented lms with compression and selective augmentation. *arXiv preprint arXiv:2310.04408*.

Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. 2024. A survey on knowledge distillation of large language models. *arXiv preprint arXiv:2402.13116*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

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, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Wenhao Yu, Hongming Zhang, Xiaoman Pan, Kaixin Ma, Hongwei Wang, and Dong Yu. 2023. Chain-of-note: Enhancing robustness in retrieval-augmented language models. *arXiv preprint arXiv:2311.09210*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

A Additional Experimental Setups

A.1 Datasets

We use publicly available multi-hop datasets. The characteristics of each dataset are as follows:

- 2WikiMultiHopQA (Ho et al., 2020): A dataset constructed using Wikipedia documents and a knowledge graph, requiring a two-hop reasoning process to answer questions.
- HotpotQA (Yang et al., 2018): A dataset where annotators created questions and answers based on multiple Wikipedia articles.
- MuSiQue (Trivedi et al., 2022b): A dataset formed by combining multiple single-hop questions into multi-hop questions requiring 2 to 4 hops.

Following the experimental setup of IRCOT (Trivedi et al., 2022a), we construct a corpus by merging the labeled documents in each dataset. We randomly sample 50,000 instances from the training data of each dataset. Since MuSiQue contains fewer than 50,000 training instances, we use its entire training set. After filtering, the final number of training samples used is 33,584 for 2WikiMultiHopQA, 30,572 for HotpotQA, and 5,515 for MuSiQue. For validation and testing, we randomly sample 500 instances from the original validation set of each dataset to construct the validation and test datasets.

A.2 Baselines

We employ the following models for our experiments. Detailed prompts for each model are provided in Section D

A.2.1 Few-shot In-Context Learning

To ensure output format consistency with other settings, we provide few-shot demonstrations.

No Retrieval The LLM generates answers without access to external documents, using few-shot exemplars and the question as a prompt.

Single-Step Retrieval The question is used as a query to search the corpus once. The top-k retrieved documents are prepended to the question as input. SuRE (Kim et al., 2024) is an advanced variant of this approach that retrieves and summarizes relevant evidence before verifying the final prediction.

Multi-Step Retrieval Multiple retrieval steps are performed according to each model's methodology to generate the final answer. For Self-Ask and ReAct, we follow the prompts provided in ITER-RETGEN.

A.2.2 Knowledge Distillation

Single-Step Retrieval SAIL (Luo et al., 2023) distills rationale generation based on informative passages retrieved from the search results. We follow the original approach using a RoBERTa entailment classification model to assess the relevance between retrieved documents and the question. Based on this relevance score, the retrieved results are formatted according to SAIL's specifications and combined with the question as input. KARD (Kang et al., 2023) is trained using data generated by prompting a teacher model with the original prompts from the KARD paper. KARD distills the teacher's reasoning while leveraging its rationale to improve retrieval. The student model is trained with the question, teacher's rationale, and documents retrieved by the rationale. CoN (Yu et al., 2023) is trained using data generated by prompting a teacher model with the original prompts from the CoN paper. The teacher produces reasoning paths that specify which documents should be referenced among the retrieved ones, and how reasoning is carried out using those documents. These paths are used to supervise the training of the student model..

Multi-Step Retrieval For Self-RAG (Asai et al., 2023), we train the student model with a teacher-generated rationale dataset, which is identical to our training dataset but further augmented with critic tokens in Self-RAG. During inference, the Self-RAG model dynamically decides whether to retrieve by generating [retrieve] critic tokens.

B Additional Experiments

B.1 Retrieval Steps Analysis

Max Retrieval Steps	3	4	5	6	7
IRCOT 70B (Teacher)	56.45	56.81	**57.23**	56.85	56.43

Table 6: Accuracy of IRCOT 70B (Teacher) by Maximum Retrieval Steps

Since the dataset consists of 2-3 hop questions, it is natural for the model to answer within 2-3 steps if retrieval works ideally. However, as the retrieval result may be incomplete, we set a higher maximum retrieval step to ensure that more relevant documents are retrieved, allowing us to better capture the necessary information. We experimented with retrieval steps ranging from 3 to 7 and found that the teacher model performed best in terms of accuracy with 5 steps as shown in Table 6.

Exact Retrieval Steps	1	2	3	4	5	Total
# Final Answers	0	11	126	206	157	500
# Correct Answers	0	4	85	135	81	305
Accuracy (%)	–	36.36	67.46	65.53	51.59	61.00

Table 7: Accuracy by Exact Number of Retrieval Steps

Upon analyzing the distribution of exact retrieval steps in StepER with a maximum retrieval step of 5, we observed a bell-shaped distribution as shown in Table 7, with 4 retrievals occurring most frequently. This suggests that the model tends to prefer 4 retrieval steps to gather sufficient evidence before answering. Although the model would ideally gather sufficient evidence within 2–3 retrieval steps, some uncertainty in earlier retrievals motivates the model to continue searching.

B.2 Retrieval Quality

Model	Accuracy	Recall	Duplicativeness
StepER-3	56.80	0.71	0.16
StepER-4	59.60	0.73	0.24
StepER-5 (Ours)	**61.00**	**0.74**	0.26
Vanilla-KD-5	54.80	0.70	0.20
Vanilla-RAG 70B-1	52.93	0.53	0.00
Vanilla-RAG 8B-1	46.15	0.53	0.00

Table 8: Comparison of Accuracy, Recall, and Duplicativeness

To analyze the quality of retrieved passages in more detail, we evaluate their quality using recall and duplicativeness. Duplicativeness is measured as the ratio of duplicated documents to the total number of retrieved documents (i.e., # overlapping docs / # retrieved docs). For each method, we report accuracy, retrieval recall (i.e., # retrieved relevant docs / # total relevant docs), and duplicativeness. The number following each method name (e.g., '3' in StepER-3) indicates the maximum number of retrieval steps used for that setting.

As shown in Table 8, StepER outperforms Vanilla-KD in recall, with its well-formed rationales effectively retrieving relevant documents, enhancing overall performance. StepER exhibits an increase in duplicativeness (26% vs. 20% in Vanilla-KD-5), with 60% of duplicated documents containing relevant passages, compared to 33% in Vanilla-KD-5. This suggests that relevant passages are retrieved multiple times, slightly raising the duplication rate.

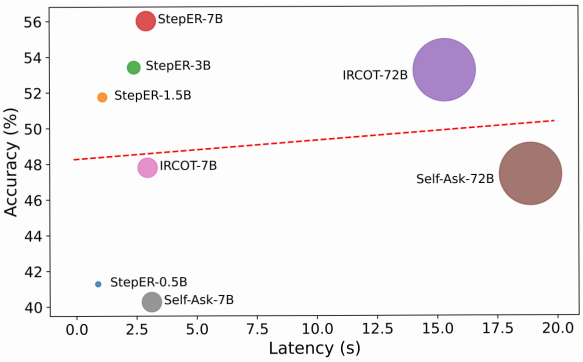


Figure 6: Accuracy (%) versus Latency (s) of StepER on Qwen2.5-Instruct. Marker size indicates model parameter count. StepER models achieve superior performance with lower latency than larger models, offering the best trade-off between efficiency and effectiveness.

B.3 Trade-off Between Latency and Accuracy

We measure the latency as the average inference time per sample on HotpotQA with Qwen2.5-Instruct models. Figure 6 illustrates the trade-off between inference latency and accuracy for different models, where the marker size indicates the model's parameter count. StepER-7B surpasses 70B-scale models in terms of accuracy, yet requires only a fraction of their latency. Thus, our evaluation confirms that StepER-7B stands out as the most efficient and effective model, delivering the best trade-off between latency and accuracy.

C Difficulty-Aware Adaptive Weighting Strategy

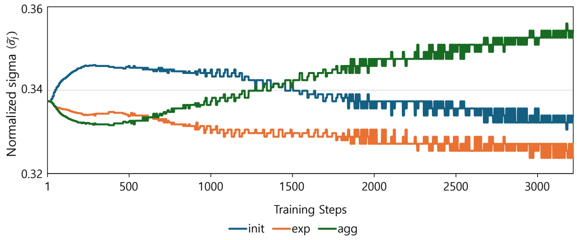


Figure 7: Evolution of normalized sigma ($\tilde{\sigma_j}$) on MuSiQue during training, reflecting changes in the relative learning difficulty of each reasoning stage: initialization (blue), expansion (orange), and aggregation (green).

To examine how the relative difficulty of each reasoning step evolves during training, we compute the normalize value of each σ_j among the set

$\{\sigma_{\text{init}}, \sigma_{\text{exp}}, \sigma_{\text{agg}}\}$, defined as:

$$\tilde{\sigma}_j = \frac{\sigma_j}{\sum_{k \in \{\text{init,exp,agg}\}} \sigma_k} \tag{5}$$

We then visualize the evolution of $\tilde{\sigma}_j$ over training in Figure 7, where $\tilde{\sigma}_{init}$, $\tilde{\sigma}_{exp}$, and $\tilde{\sigma}_{agg}$ are shown in blue, orange, and green, respectively.

Difficulty-adaptive learning is closely tied to the stepwise structure, providing justification for separating the reasoning process into distinct steps. Each step has unique characteristics, and the perceived learning difficulty for the reasoning step evolves throughout training. Thus, an adaptive approach is necessary to effectively enhance the model's capabilities for reasoning. This is further supported by the behavior of the normalized values used in difficulty-adaptive learning, which indicate shifts in task prioritization across different training steps. These adjustments ultimately contribute to improved performance as described in section 6.3. By seamlessly integrating this method with our stepwise distillation framework, we ensure that each step receives appropriate attention, reinforcing the overall effectiveness of our approach.

D Prompts

The following Tables 13 to 18 present the prompts used for our experiments

{Knowledge}
Q: Answer the following question by reasoning step-by-step.
{Question}
A:

Table 9: QA prompt for IRCOT.

{Knowledge}
Q: Answer the following question by reasoning step-by-step.
Are both Kurram Garhi and Trojkrsti located in the same country?
A: Kurram Garhi is located in the country of Pakistan. Trojkrsti is located in the country of Republic of Macedonia. Thus, they are not in the same country. So the answer is: no.

{Knowledge}
Q: Answer the following question by reasoning step-by-step.
When did the director of film Laughter In Hell die?
A: The film Laughter In Hell was directed by Edward L. Cahn. Edward L. Cahn died on August 25, 1963. So the answer is: August 25, 1963.

{Knowledge}
Q: Answer the following question by reasoning step-by-step.
What is the cause of death of Grand Duke Alexei Alexandrovich Of Russia's mother?
A: The mother of Grand Duke Alexei Alexandrovich of Russia is Maria Alexandrovna. Maria Alexandrovna died from tuberculosis. So the answer is: tuberculosis.

{Knowledge}
Q: Answer the following question by reasoning step-by-step.
Are the directors of films The Sun of the Sleepless and Nevada (1927 film) both from the same country?
A: The director of Sun of the Sleepless is Temur Babluani. The director of Nevada (1927 film) is John Waters. John Waters is from the country of America. Temur Babluani is from the country of Georgia. Thus, John Walters and Temur Babluani are not from the same country. So the answer is: no.

{Knowledge}
Q: Answer the following question by reasoning step-by-step.
When was the director of film P.S. Jerusalem born?
A: P.S. Jerusalem was directed by Danae Elon. Danae Elon was born on December 23, 1970. So the answer is: December 23, 1970.

{Knowledge}
Q: Answer the following question by reasoning step-by-step.
When did the director of film Hypocrite (Film) die?
A: The film Hypocrite was directed by Miguel Morayta. Miguel Morayta died on 19 June 2013. So the answer is: 19 June 2013.

{Knowledge}
Q: Answer the following question by reasoning step-by-step.
Where did the director of film Maddalena (1954 Film) die?
A: The film Maddalena is directed by Augusto Genina. Augusto Genina died in Rome. So the answer is: Rome.

Table 10: 7-Shot Demonstrations for IRCOT on 2WikiMultiHopQA.

{Knowledge}
Q: Answer the following question by reasoning step-by-step. Nobody Loves You was written by John Lennon and released on what album that was issued by Apple Records, and was written, recorded, and released during his 18 month separation from Yoko Ono?
A: The album issued by Apple Records, and written, recorded, and released during John Lennon's 18 month separation from Yoko Ono is Walls and Bridges. Nobody Loves You was written by John Lennon on Walls and Bridges album. So the answer is: Walls and Bridge

{Knowledge}
Q: Answer the following question by reasoning step-by-step.
When did the director of film Laughter In Hell die?
A: The film Laughter In Hell was directed by Edward L. Cahn. Edward L. Cahn died on August 25, 1963. So the answer is: August 25, 1963.

{Knowledge}
Q: Answer the following question by reasoning step-by-step. Who was born first, James D Grant, who uses the pen name of Lee Child, or Bernhard Schlink?
A: James D Grant, who uses the pen name of Lee Child, was born in 1954. Bernhard Schlink was born in 1944. Thus, Bernhard Schlink was born first. So the answer is: Bernhard Schlink.

{Knowledge}
Q: Answer the following question by reasoning step-by-step. Which band formed first, Sponge Cola or Hurricane No. 1?
A: Sponge Cola band was formed in 1998. Hurricane No. 1 was formed in 1996. Thus, Hurricane No. 1 band formed the first. So the answer is: Hurricane No. 1.

{Knowledge}
Q: Answer the following question by reasoning step-by-step. In which state of Australia will you find the themed lands Ocean parade and DreamWorks Experience both within the Dreamworld theme park complex on the Gold Coast?
A: The themed land of Ocean parade is in the state of Queensland in Australia. The themed land of The DreamWorks Experience is in the state of Queensland in Australia. Thus, both Ocean parade and The DreamWorks Experience are in the state of Queensland. So the answer is: Queensland.

{Knowledge}
Q: Answer the following question by reasoning step-by-step. Mister Magoo's Christmas Carol was produced by the same studio that produced a film that featured the only animated-film role by who?
A: Mister Magoo's Christmas Carol was produced by United Productions of America studio. United Productions of America studio produced a film Gay Purr-we, which features the voice of Judy Garland in her only animated-film role. So the answer is: Judy Garland.

{Knowledge}
Q: Answer the following question by reasoning step-by-step. How many awards did the "A Girl Like Me" singer win at the American Music Awards of 2012?
A: The singer of "A Girl Like Me" singer is Rihanna. In the American Music Awards of 2012, Rihana won one award. So the answer is: one.

Table 11: 7-Shot Demonstrations for IRCOT on HotpotQA.

{Knowledge}

Q: Answer the following question by reasoning step-by-step. What is the headquarters for the organization who sets the standards for ISO 21500?

A: The standards for ISO 21500 were set by International Organization for Standardization. The International Organization for Standardization has headquarters in Geneva. So the answer is: Geneva.

{Knowledge}

Q: Answer the following question by reasoning step-by-step. When did Britain withdraw from the country containing Hoora?

A: Hoora is in the country of Bahrain. Britain withdrew from Bahrain in 1971. So the answer is: 1971.

{Knowledge}

Q: Answer the following question by reasoning step-by-step. When did Britain withdraw from the country where the village of Wadyan is found?

A: Wadyan is in the country of Bahrain. Britain withdraw from Bahrain in 1971. So the answer is: 1971.

{Knowledge}

Q: Answer the following question by reasoning step-by-step. What shares a border with Rivière-Verte in the province WRSU-FM broadcasts in?

A: WRSU-FM was licensed to broadcast to New Brunswick. Rivière-Verte, New Brunswick shares border with Edmundston. So the answer is: Edmundston.

{Knowledge}

Q: Answer the following question by reasoning step-by-step. What genre is the record label of the performer of So Long, See You Tomorrow associated with?

A: The performer of So Long, See You Tomorrow is Bombay Bicycle Club. The record label of Bombay Bicycle Club is Island Records. The genre of Island Records is jazz. So the answer is: jazz.

{Knowledge}

Q: Answer the following question by reasoning step-by-step. What is the genre of the record label of the band that performed on the Crush Tour?

A: The Crush Tour is performed by the band Bon Jovi. The record label of Bon Jovi is Island Records. The genre of Island Records is jazz. So the answer is: jazz.

{Knowledge}

Q: Answer the following question by reasoning step-by-step. How many countries in Pacific National University's continent are recognized by the organization that mediated the truce ending the Iran-Iraq war?

A: Pacific National University is located in Khabarovsk, Russia Khabarovsk, Russian is in the continent of Asia. The entity that mediated the truce which ended the Iran-Iraq War is the UN. The number of member states that UN recognises in Asia is 53. So the answer is: 53.

Table 12: 7-Shot Demonstrations for IRCOT on MuSiQue.

Passages:
{Knowledge}
Question: {Question}
Are follow up questions needed here:

Table 13: QA prompt for Self-Ask.

```
Given the following question, answer it by providing follow up questions and intermediate answers. For each follow up question, you are given a context which is the top returned Wikipedia snippets for the question. If no follow up questions are necessary, answer the question directly.
#
Passages:
{Knowledge}
Question: Which film came out first, Blind Shaft or The Mask Of Fu Manchu?
Are follow up questions needed here: Yes.
Follow up: When did Blind Shaft come out?
Intermediate answer: Blind Shaft came out in 2003.
Follow up: When did The Mask Of Fu Manchu come out?
Intermediate answer: The Mask Of Fu Manchu came out in 1932.
So the final answer is: The Mask Of Fu Manchu
#
Passages:
{Knowledge}
Question: When did John V, Prince Of Anhalt-Zerbst's father die?
Are follow up questions needed here: Yes.
Follow up: Who is the father of John V, Prince Of Anhalt-Zerbst?
Intermediate answer: The father of John V, Prince Of Anhalt-Zerbst is Ernest I, Prince of Anhalt-Dessau.
Follow up: When did Ernest I, Prince of Anhalt-Dessau die?
Intermediate answer: Ernest I, Prince of Anhalt-Dessau died on 12 June 1516.
So the final answer is: 12 June 1516
#
Passages:
{Knowledge}
Question: Which film has the director who was born later, El Extrano Viaje or Love In Pawn?
Are follow up questions needed here: Yes.
Follow up: Who is the director of El Extrano Viaje?
Intermediate answer: The director of El Extrano Viaje is Fernando Fernan Gomez.
Follow up: Who is the director of Love in Pawn?
Intermediate answer: The director of Love in Pawn is Charles Saunders.
Follow up: When was Fernando Fernan Gomez born?
Intermediate answer: Fernando Fernan Gomez was born on 28 August 1921.
Follow up: When was Charles Saunders (director) born?
Intermediate answer: Charles Saunders was born on 8 April 1904.
So the final answer is: El Extrano Viaje
#
```

Table 14: 3-Shot Demonstrations for Self-Ask on 2WikiMultiHopQA.

```
Given the following question, answer it by providing follow up questions and intermediate answers. For each follow up question, you are given a context which is the top returned Wikipedia snippets for the question. If no follow up questions are necessary, answer the question directly.
#
Passages:
{Knowledge}
Question: What is the name of this American musician, singer, actor, comedian, and songwriter, who worked with Modern Records and born in December 5, 1932?
Are follow up questions needed here: Yes.
Follow up: Who worked with Modern Records?
Intermediate answer: Artists worked with Modern Records include Etta James, Little Richard, Joe Houston, Ike and Tina Turner and John Lee Hooker.
Follow up: Is Etta James an American musician, singer, actor, comedian, and songwriter, and was born in December 5, 1932?
Intermediate answer: Etta James was born in January 25, 1938, not December 5, 1932, so the answer is no.
Follow up: Is Little Richard an American musician, singer, actor, comedian, and songwriter, and was born in December 5, 1932?
Intermediate answer: Yes, Little Richard, born in December 5, 1932, is an American musician, singer, actor, comedian and songwriter.
So the final answer is: Little Richard
#
Passages:
{Knowledge}
Question: Between Chinua Achebe and Rachel Carson, who had more diverse jobs?
Are follow up questions needed here: Yes.
Follow up: What jobs did Chinua Achebe have?
Intermediate answer: Chinua Achebe was a Nigerian (1) novelist, (2) poet, (3) professor, and (4) critic, so Chinua Achebe had 4 jobs.
Follow up: What jobs did Rachel Carson have?
Intermediate answer: Rachel Carson was an American (1) marine biologist, (2) author, and (3) conservationist, so Rachel Carson had 3 jobs.
Follow up: Did Chinua Achebe have more jobs than Rachel Carson?
Intermediate answer: Chinua Achebe had 4 jobs, while Rachel Carson had 3 jobs. 4 is greater than 3, so yes, Chinua Achebe had more jobs.
So the final answer is: Chinua Achebe
#
Passages:
{Knowledge}
Question: Remember Me Ballin' is a CD single by Indo G that features an American rapper born in what year?
Are follow up questions needed here: Yes.
Follow up: Which American rapper is featured by Remember Me Ballin', a CD single by Indo G?
Intermediate answer: Gangsta Boo
Follow up: In which year was Gangsta Boo born?
Intermediate answer: Gangsta Boo was born in August 7, 1979, so the answer is 1979.
So the final answer is: 1979
#
```

Table 15: 3-Shot Demonstrations for Self-Ask on HotpotQA.

```
Given the following question, answer it by providing follow up questions and intermediate answers. For each follow up question, you are given a context which is the top returned Wikipedia snippets for the question. If no follow up questions are necessary, answer the question directly.
#
Passages:
{Knowledge}
Question: In which year did the publisher of In Cold Blood form?
Are follow up questions needed here: Yes.
Follow up: What business published In Cold Blood?
Intermediate answer: In Cold Blood was published in book form by Random House.
Follow up: Which year witnessed the formation of Random House?
Intermediate answer: Random House was form in 2001.
So the final answer is: 2001
#
Passages:
{Knowledge}
Question: Who was in charge of the city where The Killing of a Sacred Deer was filmed?
Are follow up questions needed here: Yes.
Follow up: In which city was The Killing of a Sacred Deer filmed?
Intermediate answer: The Killing of a Sacred Deer was filmed in Cincinnati.
Follow up: Who was in charge of Cincinnati?
Intermediate answer: The present Mayor of Cincinnati is John Cranley, so John Cranley is in charge.
So the final answer is: John Cranley
#
Passages:
{Knowledge}
Question: Where on the Avalon Peninsula is the city that Signal Hill overlooks?
Are follow up questions needed here: Yes.
Follow up: What city does Signal Hill overlook?
Intermediate answer: Signal Hill is a hill which overlooks the city of St. John's.
Follow up: Where on the Avalon Peninsula is St. John's located?
Intermediate answer: St. John's is located on the eastern tip of the Avalon Peninsula.
So the final answer is: eastern tip
```

Table 16: 3-Shot Demonstrations for Self-Ask on MuSiQue.

You will be given a reasoning task with passage(s), a question, gold answer(s), and generated answer from model.
Your task is to evaluate the generated answer as either 0 or 1 based on the following criteria.
Consider the passages when making your evaluation.
You must answer the evaluation form using json format.

Evaluation Criteria:
1. Reasoning Initialization: Evaluate how well the generated answer starts the reasoning path based on the given passages and question. Does the first sentence provide a logical and relevant foundation for the rest of the reasoning? Consider the following:
- If the first reasoning step provides a necessary foundation for expanding the reasoning, evaluate it positively.
- If the first reasoning path is irrelevant or diverges from addressing the question directly, evaluate it negatively regardless of whether the answer is correct or incorrect.
2. Reasoning Expansion: Assess how well the generated answer extracts and applies relevant information from the passages to address the question. Does each subsequent sentence logically expand upon the first sentence to develop the reasoning effectively? Consider the following:
- If the model correctly extracts key information and logically expands upon it to support the reasoning, evaluate positively.
- If relevant information exists in the passages but is ignored or misused, evaluate negatively.
3. Reasoning Aggregation: Assess the alignment between the reasoning path and the final answer. Does the reasoning path logically lead to the final answer and ensure its correctness based on the provided reasoning? Consider the following:
- If both the reasoning path and the final answer are logically consistent, correct, and directly address the question, evaluate it positively.
- If the reasoning path contains correct intermediate steps but the final answer is logically inconsistent or incorrect, evaluate it negatively.
- If the reasoning path is incorrect but the final answer happens to be correct, also evaluate it negatively.

Evaluation Form:
- Reasoning Initialization: {{0 / 1}}
- Reasoning Expansion: {{0 / 1}}
- Reasoning Aggregation: {{0 / 1}}

Question:
{question}
Gold Answer List:
{gold_answer_list}
Passages:
{passage}
Generated Answer:
{generated_answer}

Table 17: GPT evaluation prompt for assessing reasoning abilities

You will be given a reasoning task with a question, a gold answer, a model-generated rationale, and two sub-questions with their gold answers.
Your task is to evaluate whether the model-generated rationale provides enough information to answer the two sub-questions and whether the answers are correct.
Please read and understand these instructions carefully before proceeding with the evaluation.
Refer back to them as needed during evaluation.
You must answer the evaluation form using json format.

Evaluation Criteria:
1. Sub-question Answerability
- Evaluate whether each sub-question can be correctly answered using only the given rationale.
- DO NOT use external knowledge beyond the rationale.
- If both sub-questions can be correctly answered using only the rationale, evaluate it as 1.0.
- If only one sub-question can be correctly answered, evaluate it as 0.5.
- If neither sub-question can be answered, evaluate it as 0.0.

2. Answer Correctness
- Evaluate whether the answers to the main question and the two sub-questions are correct.
- Compare each model-generated answer with its corresponding gold answer.
- If the model-generated answer is correct, mark it as "correct"; otherwise, mark it as "wrong".
- Provide the correctness evaluation in the form of a list:
- First element: Whether the model-generated answer to the main question is correct.
- Second element: Whether the Sub-Question 1 can be correctly answered using only the model-generated rationale.
- Third element: Whether the Sub-Question 2 can be correctly answered using only the model-generated rationale.

Evaluation Form:
- Sub-question Answerability: {{1.0 / 0.5 / 0.0}}
- Answer Correctness: ["{{correct / wrong}}", "{{correct / wrong}}", "{{correct / wrong}}"]

Input:
- Main Question: {question}
- Gold Answer for Main question: {answer}
- Model-Generated Rationale: {rationale}
- Sub-Question 1: {sub_question_1}
- Gold Answer for Sub-Question 1: {sub_answer_1}
- Sub-Question 2: {sub_question_2}
- Gold Answer for Sub-Question 2: {sub_answer_2}

Table 18: GPT evaluation prompt for assessing rationale validity on SubQA