SimRAG: Self-Improving Retrieval-Augmented Generation for Adapting Large Language Models to Specialized Domains

Ran Xu[1,2*], **Hui Liu**[2], **Sreyashi Nag**[2], **Zhenwei Dai**[2], **Yaochen Xie**[2], **Xianfeng Tang**[2], **Chen Luo**[2], **Yang Li**[2], **Joyce C. Ho**[1], **Carl Yang**[1], **Qi He**[2]
[1] Emory University [2] Amazon
{ran.xu,joyce.c.ho,j.carlyang}@emory.edu, liunhu@amazon.com

Abstract

Retrieval-augmented generation (RAG) enhances the question answering (QA) abilities of large language models (LLMs) by integrating external knowledge. However, adapting general-purpose RAG systems to specialized fields such as science and medicine poses unique challenges due to distribution shifts and limited access to domain-specific data. To tackle this, we propose SimRAG, a self-training approach that equips the LLM with joint capabilities of question answering and question generation for domain adaptation. Our method first fine-tunes the LLM on instruction-following, question-answering, and search-related data. Then, it prompts the same LLM to generate diverse domain-relevant questions from unlabeled corpora, with an additional filtering strategy to retain high-quality synthetic examples. By leveraging these self-generated synthetic examples, the LLM can improve their performance on domain-specific RAG tasks. Experiments on 11 datasets, spanning two backbone sizes and three domains, demonstrate that SimRAG outperforms baselines by 1.2%–8.6%.

1 Introduction

Retrieval-augmented generation (RAG) (Lewis et al., 2020; Gao et al., 2023; Gutiérrez et al., 2024; Asai et al., 2024) is a powerful technique that enhances large language models (LLMs) for various knowledge-intensive tasks such as question answering (QA) by incorporating external knowledge sources. This method not only customizes responses to handle long-tail knowledge but also avoids the need for costly model retraining (Ovadia et al., 2023). Additionally, RAG helps reduce the issue of LLM hallucination by ensuring responses are grounded in relevant evidence (Shuster et al., 2021), thereby improving the overall accuracy and reliability of LLM outputs.

While extensive research has focused on developing effective (Asai et al., 2024; Lin et al., 2024; Liu et al., 2024) and efficient (Xu et al., 2024a) RAG systems for general-domain QA tasks, adapting RAG to specialized domains for LLMs poses significant challenges. These models often struggle with distribution shifts and fail to accurately extract information from domain-specific contexts (Miller et al., 2020; Liu et al., 2022). Moreover, directly using black-box LLMs (OpenAI, 2023; Anthropic, 2023; Wang et al., 2023b) in specialized domains raises concerns about privacy when dealing with sensitive proprietary data. It is essential to fine-tune LLMs on domain-relevant QA tasks to unlock the full potential of LLM-based RAG systems in specialized domains.

Despite the critical need for domain-specific fine-tuning, the primary challenge lies in the acquisition of high-quality fine-tuning data towards RAG applications. Prior works rely on continuous pretraining (Chen et al., 2023; Zhang et al., 2024a) on specialized corpora or fine-tuning on domain-specific instruction-tuning data (Wu et al., 2024; Wadden et al., 2024). However, the mismatch between these general-purpose tasks and domain-specific QA hinders their effectiveness. More recently, several approaches (Liu et al., 2024; Schimanski et al., 2024; Zhang et al., 2024c) use synthetic data from powerful LLMs (e.g., GPT-4) to create QA fine-tuning datasets. While promising, these methods are costly, inefficient, and lack explicit quality control over the generated outputs. Additionally, the direct use of proprietary corpora with black-box LLMs introduces privacy concerns, making these methods unsuitable for sensitive domains.

To tackle the data scarcity issue mentioned above, we propose SimRAG[1], a self-improving approach to harness the LLMs' own capabilities to generate pseudo-labeled data for domain adap-

[*] Work done during an internship at Amazon.

[1] Self-improving Retrieval-Augmented Generation.

tative question answering. Our method is inspired by the success of self-training in LLM development, where models are refined using synthetic examples generated from unlabeled corpora (Wang et al., 2022; Li et al., 2024). However, for RAG applications, special considerations are needed to adapt LLMs for generating questions that require external context to answer. The core objective of SimRAG is to fine-tune a single LLM to perform two complementary tasks: *question answering with context* and *question generation from context*. Both tasks involve extracting and summarizing relevant information from the context, allowing them to mutually reinforce each other.

Specifically, we design a two-stage procedure to adapt LLMs for domain QA, we first fine-tune LLMs on *instruction-following*, *question answering*, and *question generation* data from general-domains. This step equips LLMs with basic instruction-following and context utilization skills. Then, to specialize the model for domain-specific tasks, we then harness unlabeled domain corpora, prompting the same LLM to generate high-quality QA pairs grounded in the context of these specialized domains. To further *enhance the quality* of synthetic pairs, we incorporate multiple task types to improve the model's generalization capabilities, combined with round-trip consistency filtering technique (Bartolo et al., 2021) to preserve generated QA pairs only when the original context is retrieved among top results. With these pseudo-labeled (*question*, *passage*, *answer*) tuples generated by LLMs, we continuously fine-tune the models with those synthetic examples. This pipeline allows the LLM to progressively refine its output on synthetic pairs, thus adapting itself towards domain-specific QA applications.

We conduct experiments on three different domains spanning from biomedical, natural/social sciences, and computer science (CS), where we observe SimRAG consistently achieve better performance than other domain-specific LLMs and general-domain retrieval-augmented LMs. Qualitative studies highlight the benefits of joint training in question answering and generation, along with diverse, denoised QA pairs.

Our contribution can be summarized as follows:

- We propose SimRAG, a RAG framework that enhances LLM's capability for question answering on specialized domains.
- We design a novel instruction fine-tuning approach that enables LLMs to perform both question answering and question generation. This joint capability facilitates self-improvement through self-training on generated synthetic data, leading to enhanced model performance.
- We validate our approach with empirical studies across 11 datasets from three distinct domains, demonstrating that SimRAG outperforms baseline models by 1.2%–8.6%.

2 Related Work

Retrieval-augmented generation. RAG has emerged as a powerful tool in knowledge-intensive NLP tasks such as language modeling (Borgeaud et al., 2022) and question answering (Lewis et al., 2020; Shi et al., 2024a). The typical approach involves integrating a retriever with the LLM generator and designing a fine-tuning process to align the retriever with LLM capabilities. To further refine RAG, recent research explored various enhancements. These include developing dynamical retrieval processes to refine the relevance of fetched content (Jiang et al., 2023; Jeong et al., 2024; Su et al., 2024), and filtering out irrelevant contexts to robustify RAG (Yoran et al., 2024; Yu et al., 2024, 2023; Wang et al., 2024). Additionally, several studies have developed instruction-tuning methods aimed specifically at improving search and RAG capabilities of LLMs (Liu et al., 2024; Lin et al., 2024; Dong et al., 2024; Wei et al., 2024).

Self-training. Self-training (or Pseudo-Labeling) is one of the earliest approaches to semi-supervised learning (Rosenberg et al., 2005). The method uses a teacher model to generate new labels on which a student model is fitted. Self-training has been widely adopted for various NLP tasks including text classification (Du et al., 2021), natural language understanding (Vu et al., 2021) and ranking (Wang et al., 2022). Recently, the idea of self-training has also been applied to LLM instruction fine-tuning (Yuan et al., 2024; Li et al., 2024), reasoning (Pang et al., 2024), and alignment (Gulcehre et al., 2023), yet to the best of our knowledge, this pipeline has not been widely explored for RAG applications. The major drawback of self-training is that it is vulnerable to label noise (Arazo et al., 2020). There are several approaches to stabilize the self-training, with sample selection (Li et al., 2024) and reweighting (Wang et al., 2021) strategies.

Domain-specific LLMs. Most domain-specific LLMs rely on continuous pretraining (Labrak

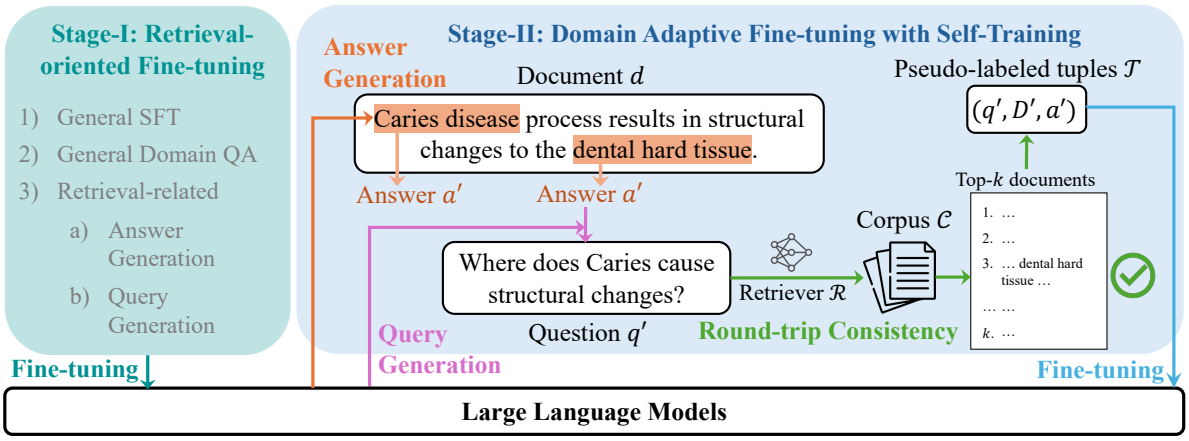

Figure 1: Two-stage fine-tuning framework for our proposed method SimRAG. The model is first fine-tuned on retrieval-related data. Then, it generates pseudo-labeled tuples by first extracting candidate answers from the corpus, and then generating candidate questions conditioned on both document and answer. The LLM is further fine-tuned on pseudo-labeled examples filtered with round-trip consistency.

et al., 2024; Chen et al., 2023; Xu et al., 2024b) or domain-specific fine-tuning (Wu et al., 2024; Zhang et al., 2024a, 2023; Wadden et al., 2024; Shi et al., 2024b), with little focus on adapting models for domain-specific RAG settings. Relevant works (Zhang et al., 2024c; Schimanski et al., 2024) use strong GPT models for synthetic data generation in RAG scenarios. In contrast, SimRAG leverages the same LLM for both question generation and answering, enabling self-improvement and offering a more cost-effective approach for adapting LLMs to domain-specific QA tasks.

3 Methodology

3.1 Problem Setup

In a RAG problem, we aim to generate answers for queries based on a set of supporting documents or contexts. Specifically, for a query q, an retriever $\mathcal{R}$ is utilized to retrieve top-k most relevant contexts $\mathcal{D} = \{d_1, d_2, ..., d_k\}$ from a large corpus $\mathcal{C}$. The LLM $\mathcal{M}_\theta$ then generates an answer a to the query q based on the retrieved context $\mathcal{D}$.

In this work, we aim to improve the LLM's QA capability in RAG system towards specialized domains where only unlabeled corpus $\mathcal{C}$ is available. As shown in Figure 1, our approach first learns from retrieval-oriented instruction data in the general domain in Stage-I and then augments $\mathcal{T}$ with pseudo-labeled $\mathcal{T}' = (q', \mathcal{D}', a')$ tuples in Stage-II, where $\mathcal{D}'$ is sampled from the specialized domain $\mathcal{C}$ for self-training. The overall objective of our study is to adapt the LLM $\mathcal{M}_\theta$ to specialized domains with $\mathcal{T} \cup \mathcal{T}'$.

3.2 Stage-I: Retrieval-oriented fine-tuning

To start with, we leverage instruction fine-tuned LLMs as the backbone (e.g. `meta-llama/Meta-Llama-3-8B-Instruct`). Although these models have been instruction finetuned, they still exhibit a deficiency in leveraging context information to answer domain-specific questions. To improve their abilities on knowledge-intensive tasks, we fine-tune the LLM with retrieval-oriented tasks. Specifically, we follow Lin et al. (2024); Liu et al. (2024) and leverage the training data blend that consists of the following components:

(1) **General Instruction Fine-tuning (SFT) data.** To help maintain the models' ability to comprehend and follow instructions, we leverage the SFT data including OpenAssistant (Köpf et al., 2023), Dolly (Conover et al., 2023), SODA (Kim et al., 2023), ELI5 (Fan et al., 2019), Self-Instruct (Wang et al., 2023a), and Unnatural Instructions (Honovich et al., 2022). Note that we make sure there is no overlap between SFT data and test data from target tasks.

(2) **General domain Context-aware QA data.** To bolster the LLMs' general RAG skills of generating accurate answers grounded in relevant contexts, we fine-tune them on a diverse array of general domain question-answering datasets. This includes DROP (Dua et al., 2019), NQ (Kwiatkowski et al., 2019), Squad (Rajpurkar et al., 2016), NarrativeQA (Kočiský et al., 2018), Quoref (Dasigi et al., 2019), ROPES (Lin et al., 2019), OpenbookQA (Mihaylov et al., 2018), LogiQA (Liu et al., 2020), TAT-QA (Zhu et al., 2021), We-

bGLM (Liu et al., 2023), StrategyQA (Geva et al., 2021), BoolQ (Clark et al., 2019), FaVIQ (Park et al., 2022) and FEVER (Thorne et al., 2018) datasets, where for each sample, a query q and its relevant context $\mathcal{D}$ is given, and the LLM is trained to generate answer a to the query.

(3) **General Retrieval-related Data:** To better generate high-quality pseudo-labeled QA samples in the next stage, we incorporate retrieval-related data to improve two specific skills of LLMs: (a) *Answer Generation*: where a grounding document is given, and the LLMs are trained to generate candidate spans from the context that are likely to be answers to some questions. In this part, we incorporate Squad 1.1 and 2.0 versions (Rajpurkar et al., 2016), DROP (Dua et al., 2019) and WebQuestions (Berant et al., 2013) datasets. (b) *Query Generation*: where an answer and its grounding document are given, and the LLMs are trained to generate a query based on the document and answer. In this part, we leverage NQ (Kwiatkowski et al., 2019), Squad 1.1 (Rajpurkar et al., 2016), StrategyQA (Geva et al., 2021), WebQuestions (Berant et al., 2013), FaVIQ (Park et al., 2022) and FEVER (Thorne et al., 2018) datasets.

The details for each dataset (e.g. the instruct format and the amount of data used) are deferred to Appendix A. For each sample in the fine-tuning dataset, we adopt a standard instruction finetuning objective, computing the loss exclusively on the tokens of the assistant's response.

3.3 Stage-II: Domain Adaptive Fine-tuning

The model after Stage-I is only trained in the general domains. When directly adopting them to specialized applications, the performance can still be suboptimal due to the distribution shift issue (Miller et al., 2020). To tailor the LLMs for specialized domains and address the scarcity of labeled data in these areas, we employ a self-training approach leveraging domain-specific unlabeled corpora. This method capitalizes on the model's enhanced capabilities from the previous retrieval-augmented fine-tuning stage. We utilize the fine-tuned LLM to generate pseudo-labeled training samples $\mathcal{T}' = (q', \mathcal{D}', a')$ by creating queries grounded in the unlabeled text and gathering the corresponding retrieved documents.

Specifically, we conduct a two-step procedure to synthesize additional training data, which corresponds to the two skills learned in Stage-I: (a) *Answer Generation*: for each document $d_i \in \mathcal{C}$, where $\mathcal{C}$ is the unlabeled corpus, we prompt our fine-tuned LLMs to generate several candidate spans $a_i^1, a_i^2, \ldots, a_i^m$ that are likely to be answers to some questions. Formally, the model generates $a_i^j \sim p_\theta(\cdot|d_i)$ for $j = 1, \ldots, m$. (b) *Answer-conditioned Query Generation*: for each candidate answer a_i^j and its corresponding document d_i, we prompt the fine-tuned LLM again to generate candidate questions $q_i^j \sim p_\theta(\cdot|a_i^j, d_i)$, with a_i^j as the ground truth answer and d_i as the supporting context. This gives us the pseudo-labeled query-answer pair (q_i^j, a_i^j) based on the context d_i.

During this process, we adopt two additional strategies, namely *diverse question generation* and *data filtering*, to further improve the quality of the synthetic pairs. For diverse question generation, we prompt the LLM to create various types of questions, including *short-span question-answering*, *multiple-choice question-answering*, and *claim verification* tasks. While short-span questions follow the same pipeline as previously described, multiple-choice questions are constructed by using alternative candidate answers from the same unlabeled corpus in step (a) as incorrect options. Claim verification, on the other hand, bypasses the answer generation step; instead, the LLM generates a claim that can be either supported or refuted by the provided document. By injecting different question types, we prevent the LLM from overfitting to a specific output format and improve the model's generalization ability across different QA tasks.

After generating large amounts of candidate QA pairs, we implement a filtering step to keep only high-quality QA pairs. We define high-quality QA pairs as those that are answerable using the top-k retrieved contexts. Specifically, we retain only those samples where the ground truth answer a_i' is present in the top-k documents retrieved by a strong retriever, such as Dragon (Lin et al., 2023), based on the generated query q_i'. Formally, the sample is retained if $a_i' \in \mathcal{D}_i'^k$, where $\mathcal{D}_i'^k$ denotes the top-k documents retrieved for query q_i'. From these retained samples, we create pseudo-labeled training tuples $\mathcal{T}' = (q_i', \mathcal{D}_i', a_i')_{i=1}^n$.

With the created synthetic tuples $\mathcal{T}'$, we augment it with the SFT data $\mathcal{T}_{\text{SFT}}$ and the general domain context-aware QA data from Stage-I $\mathcal{T}_{\text{gen}}$, to continuously fine-tune our models, enhancing the LLMs' QA abilities within the specific domain. The size and blending ratio of the pseudo-labeled samples can be found in Appendix A.

4 Experimental Setup

4.1 Tasks and Datasets

We evaluate our model across a total of 11 datasets spanning the *medical*, *scientific* and *computer science* domains. For the medical domain, we include the five datasets in the MIRAGE benchmark (Xiong et al., 2024), including PubMedQA (Jin et al., 2019), BioASQ (Tsatsaronis et al., 2015), MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), the medical subsets in MMLU (Hendrycks et al., 2021), and two additional open-ended QA datasets LiveQA (Abacha et al., 2017), and MedicationQA (Abacha et al., 2019). For the scientific domain, we consider ARC-challenge (Clark et al., 2018), SciQ (Welbl et al., 2017)[2], and the scientific subsets (14 subtasks in total) in MMLU (Hendrycks et al., 2021). For computer science, we use CS-Bench (Song et al., 2024) for evaluation. We distinguish the computer science domain from the broader scientific domain as the scientific domain predominantly covers natural and social sciences, with limited representation of computer science topics. We use accuracy as the evaluation metric for multiple-choice and True-or-False questions, Rouge-L and MAUVE for open-ended questions, Exact Match (EM) and F1 for Fill-in-the-blank questions, with Rouge-L and F1 as the main metrics, respectively. An exception is CS-Bench, where we follow the original paper's evaluation method by using GPT-4 as a judge for fill-in-the-blank and open-ended questions.

For the medical domain, we use the corpora from Textbooks (Jin et al., 2021), Wikipedia and PubMed articles[3] to generate pseudo-labeled samples in Stage-II. For the scientific domain, we leverage Wikipedia. For the CS domain, we use Wikipedia CS Subset[4] and arXiv articles[5].

4.2 Baselines

We categorize our baselines into four groups: (1) *Off-the-shelf general domain LLMs*, which include GPT-3.5 (OpenAI, 2022), GPT-4 (OpenAI, 2023), Llama3-8B-it (Meta-AI, 2024), and Gemma2-27B-it (Team et al., 2024). (2) *Off-the-shelf domain-specific LLMs*, including PMC-llama-13B (Wu et al., 2024), MEDITRON-70B (Chen et al., 2023), AdaptLLM-v2-8B (Cheng et al., 2024), BioMistral-7B (Labrak et al., 2024) and MedLlama3-8B (John Snow Labs, 2024) in the medical domain, as well as SciTulu 7B and 70B (Wadden et al., 2024) in both the scientific domain and the computer science domain, due to the absence of LLMs specifically fine-tuned for the computer science domain. (3) *General domain retrieval-augmented LLMs*, which include Self-RAG-13B (Asai et al., 2024), ChatQA1.5-8B and 70B (Liu et al., 2024). (4) *Domain-specific Retrieval-augmented LLMs*, including RAFT (Zhang et al., 2024c) and EvidenceRAG[6] (Schimanski et al., 2024). Since RAFT and EvidenceRAG have not released their checkpoints, we re-implemented their methods using the same backbones as our approach. Note that for all the baseline models, we conduct the zero-shot evaluation and augment the context with retrieval for fair comparison. We also note that we do not compare with several domain-specific baselines such as (Zhang et al., 2024b; Nori et al., 2023) which have access to task-specific examples that overlap with our evaluation tasks.

4.3 Implementation Details

We use Llama3-it 8B (Meta-AI, 2024) and Gemma2-it 27B (Team et al., 2024) as our backbones. For the Gemma-2 model, we use LoRA (Hu et al., 2022) ($r = 32, \alpha = 32$) during fine-tuning due to resource constraints. For both stages, we set the global batch size to 64, with gradient accumulation as 8 and train the model for 1 epoch. For Stage-I, the learning rate is set to $5e-7$ and for Stage-II, it is set to $2e-7$ for the Llama3 backbone and $5e-7$ for the Gemma backbone. AdamW optimizer (Loshchilov and Hutter, 2019) is used with $\beta_1 = 0.9$ and $\beta_2 = 0.95$. To create context-enhanced examples for our synthetic queries, we use Dragon (Lin et al., 2023) to extend context length for SimRAG and baselines, which improves RAG model robustness (Yu et al., 2024, 2023). For retrieval during evaluation on medical datasets, we follow the original MIRAGE benchmark by using the top-10 retrieval results as context, ensembled from multiple models. For other datasets, we fetch the top-10 passages by Google Search[7]. All experiments are conducted on 8 NVIDIA A100 GPUs. The prompt format for answer and question gener-

[2]We convert the multiple-choice questions in SciQ into short-phrase answer generation tasks to better assess the model's generative capabilities.

[3]https://pubmed.ncbi.nlm.nih.gov/

[4]https://huggingface.co/datasets/AlaaElhilo/Wikipedia_ComputerScience

[5]https://huggingface.co/datasets/CCRss/arxiv_papers_cs

[6]We named this method ourselves, as the model does not have an officially designated name.

[7]https://www.searchapi.io

Table 1: Results of our proposed method and baselines in the medical domain. All the presented methods use RAG for inference. **Bold** and underline highlight the best and second best performance, respectively. *: the main metric used for average calculation. †: models trained using synthetic data generated from GPT-4. ‡: our own implementation of the models with the same unlabeled corpora. The notations are the same for the following tables.

Datasets	PubMedQA	BioASQ	MedQA	MedMCQA	MMLU-med	LiveQA	MedicationQA	Avg.
Metrics	ACC	ACC	ACC	ACC	ACC	Rouge-L* / MAUVE	Rouge-L* / MAUVE	—
Proprietary LLMs, For Reference Only								
GPT-3.5 (OpenAI, 2022)	67.40	90.29	66.61	58.04	75.48	42.3 / 62.5	36.3 / 46.0	62.35
GPT-4 (OpenAI, 2023)	70.60	92.56	82.80	66.65	87.24	44.0 / 65.9	41.5 / 59.2	**69.34**
Medical LLMs								
PMC-Llama 13B (Wu et al., 2024)	56.00	65.21	42.58	48.29	52.53	35.7 / 60.6	36.4 / 38.3	48.10
MEDITRON 70B (Chen et al., 2023)	56.40	76.86	49.57	52.67	65.38	—	—	—
AdaptLLM-v2 8B (Cheng et al., 2024)	45.00	78.80	43.13	42.74	51.24	30.2 / 48.0	39.2 / 51.4	47.19
BioMistral 7B (Labrak et al., 2024)	59.20	82.69	32.52	32.20	47.47	43.1 / 63.2	39.6 / 51.9	48.11
MedLlama3 8B (John Snow Labs, 2024)	74.20	83.50	61.43	61.18	77.13	27.9 / 45.2	29.8 / 35.0	59.31
Retrieval-Augmented LLMs								
Self-RAG 13B† (Asai et al., 2024)	71.20	73.70	48.60	44.00	53.90	35.6 / 54.1	39.3 / 46.4	52.33
ChatQA1.5 8B (Liu et al., 2024)	66.40	82.69	42.36	46.97	61.40	39.3 / 65.5	39.9 / 48.9	54.15
ChatQA1.5 70B (Liu et al., 2024)	74.80	83.17	68.89	62.54	80.51	40.1 / 66.3	40.8 / 50.2	64.40
‡*Backbone: Llama3-8B-Instruct*								
Llama3-8B-it (Meta-AI, 2024)	64.60	88.51	55.30	58.91	69.79	34.1 / 54.1	37.2 / 45.6	58.34
RAFT 8B† (Zhang et al., 2024c)	73.40	88.67	54.28	60.15	70.25	36.2 / 55.6	38.9 / 56.4	60.26
EvidenceRAG 8B† (Schimanski et al., 2024)	75.00	90.61	57.74	61.13	72.27	36.6 / 57.8	34.6 / 53.6	61.14
SimRAG 8B	**80.00**	91.75	62.92	**67.51**	75.57	**44.4** / 66.6	40.1 / 57.4	**66.04**
w/o Stage II	78.00	90.45	60.56	65.22	74.56	42.8 / 62.9	38.5 / 55.6	64.30
‡*Backbone: Gemma2-27B-Instruct*								
Gemma2-27B-it (Team et al., 2024)	56.20	89.32	59.70	57.30	75.67	37.4 / 52.8	40.2 / 57.0	59.40
RAFT 27B† (Zhang et al., 2024c)	67.20	91.70	62.22	61.56	78.97	39.4 / 62.2	40.2 / 48.2	63.04
EvidenceRAG 27B† (Schimanski et al., 2024)	63.00	90.61	62.14	61.80	79.43	34.5 / 58.6	34.5 / 44.6	60.85
SimRAG 27B	73.60	**92.07**	**63.63**	64.16	**81.63**	39.9 / **66.8**	**41.2** / **62.1**	65.17
w/o Stage II	66.00	91.59	62.45	58.67	79.61	37.2 / 61.6	40.8 / 58.6	62.33

Table 2: Results of our proposed method and baselines in the scientific domain.

Models	MMLU-sci	ARC	SciQ	Avg.
Metrics	ACC	ACC	EM / F1*	—
Proprietary LLMs, For Reference Only				
GPT-3.5 (OpenAI, 2022)	66.40	75.30	40.30 / 62.73	68.14
GPT-4 (OpenAI, 2023)	87.46	94.03	43.24 / 66.03	**82.51**
Scientific LLMs				
SciTulu 7B (Wadden et al., 2024)	55.95	53.84	22.2 / 40.55	50.11
SciTulu 70B (Wadden et al., 2024)	71.80	52.82	18.6 / 36.69	53.77
Retrieval-Augmented LLMs				
Self-RAG 13B† (Asai et al., 2024)	48.69	73.10	31.60 / 51.87	57.89
ChatQA 8B (Liu et al., 2024)	54.46	52.22	40.40 / 60.60	55.76
ChatQA 70B (Liu et al., 2024)	75.21	81.06	50.00 / 68.41	74.89
‡*Backbone: Llama3-8B-Instruct*				
Llama3-8B-it (Meta-AI, 2024)	67.15	71.08	20.80 / 42.47	60.23
RAFT 8B† (Zhang et al., 2024c)	69.22	73.12	48.20 / 68.56	70.30
EvidenceRAG 8B† (Schimanski et al., 2024)	71.59	75.34	53.10 / 70.11	72.35
SimRAG 8B	77.31	81.40	57.50 / 72.17	76.96
w/o Stage II	75.95	80.20	53.80 / 70.16	75.44
‡*Backbone: Gemma2-27B-Instruct*				
Gemma2-27B-it (Team et al., 2024)	76.11	85.75	44.80 / 66.99	76.28
RAFT 27B† (Zhang et al., 2024c)	78.79	86.95	53.10 / 70.91	78.88
EvidenceRAG 27B† (Schimanski et al., 2024)	78.84	86.69	45.60 / 67.50	77.68
SimRAG 27B	**81.28**	**88.65**	**58.10 / 74.99**	**81.64**
w/o Stage II	78.38	86.86	54.50 / 72.00	79.08

Table 3: Results of our proposed method and baselines in the computer science domain. MC, AS, FB, OG stands for multiple-choice, assertion, fill-in-the-blank and Open-ended generation, respectively.

Models	MC	AS	FB	OE	Overall
Metrics	ACC	ACC	Auto	Auto	—
Proprietary LLMs, For Reference Only					
GPT-3.5 (OpenAI, 2022)	54.89	67.30	42.93	50.11	55.74
GPT-4 (OpenAI, 2023)	71.48	73.62	56.87	71.43	**70.34**
Scientific LLMs					
SciTulu 7B (Wadden et al., 2024)	38.40	56.56	27.66	32.29	40.44
SciTulu 70B (Wadden et al., 2024)	44.24	60.18	31.06	54.76	46.87
Retrieval-Augmented LLMs					
Self-RAG 13B† (Asai et al., 2024)	29.87	54.52	30.64	24.94	34.56
ChatQA 8B (Liu et al., 2024)	35.33	60.18	27.66	29.82	39.11
ChatQA 70B (Liu et al., 2024)	54.94	62.67	34.89	38.53	53.07
‡*Backbone: Llama3-8B-Instruct*					
Llama3-8B-it (Meta-AI, 2024)	52.69	60.41	26.81	44.12	50.80
RAFT 8B† (Zhang et al., 2024c)	54.57	60.86	32.76	40.23	52.38
EvidenceRAG 8B† (Schimanski et al., 2024)	54.42	62.67	35.02	42.30	53.06
SimRAG 8B	60.63	64.93	34.47	47.11	57.63
w/o Stage II	59.88	61.99	34.47	46.82	56.55
‡*Backbone: Gemma2-27B-Instruct*					
Gemma2-27B-it (Team et al., 2024)	59.96	62.22	40.00	**57.50**	58.08
RAFT 27B† (Zhang et al., 2024c)	60.93	66.06	39.15	53.80	59.07
EvidenceRAG 27B† (Schimanski et al., 2024)	60.63	62.22	40.85	54.40	58.34
SimRAG 27B	**62.87**	**66.74**	**43.83**	54.60	**60.96**
w/o Stage II	61.00	65.84	41.70	54.00	59.36

ation and inference can be found in Appendix E.

5 Experimental Results

5.1 Main Results

Table 1, Table 2, and Table 3 present the experimental results for the medical, scientific, and computer science domains, respectively. The results of the 14 tasks in MMLU-sci can be found in Appendix D From the results, we have the following findings:

(1) SimRAG consistently outperforms baselines across these domains and a variety of question-answering formats. In medical, scientific, and computer science domain, the average performance gain is 8.01%, 6.37%, 8.61% over the Llama variant and 1.19%, 3.50%, 3.20% over the Gemma variant, respectively. Besides, SimRAG also achieves

Table 4: Performance of SimRAG using Llama-3-8b-it as the backbone and its variants across medical datasets.

Method	PubMedQA	BioASQ	MedQA	MedMCQA	MMLU-med	LiveQA	MedicationQA	Avg.
SimRAG 8B	80.00	91.75	62.92	67.51	75.57	44.4	40.1	66.04
SimRAG w/o general SFT data	79.60	90.78	59.47	61.92	73.09	39.9	38.9	63.38
SimRAG w/o general-domain QA	80.20	91.10	61.04	65.31	72.91	42.8	39.4	64.68
SimRAG w/o general retrieval data	79.40	90.94	57.97	62.42	71.72	39.4	38.9	62.96
SimRAG w/o Stage-I	76.80	89.81	57.97	60.02	70.71	39.3	38.5	61.87

comparable performance to strong proprietary models: when using Gemma2-27B as the backbone, we achieve 93.99%, 98.95% and 86.66% of the performance of GPT-4. This demonstrates the effectiveness and robustness of SimRAG in adapting general-domain LLMs to specialized domain knowledge using only unlabeled corpora.

(2) *Domain-specific LLMs* (e.g. SciTulu and MedLlama), although fine-tuned on relevant data, underperform compared to SimRAG because they are not optimized for RAG tasks, where effectively utilizing retrieved context is crucial. As a result, they struggle to incorporate relevant context into their answers, leading to weaker performance. On the other hand, *general-domain RAG models* (e.g. ChatQA) face distribution shifts when applied to specialized tasks, as they struggle to integrate the retrieved domain-specific knowledge accurately.

(3) *Domain-specific retrieval-augmented LLMs* such as RAFT and EvidenceRAG still show suboptimal performance despite utilizing the powerful (yet expensive) GPT-4 model to generate synthetic training data. In contrast, SimRAG, fine-tuned specifically for the QA generation task, produces more accurate and contextually relevant synthetic QA pairs, leading to better downstream performance across all QA tasks.

(4) Although the CS domain is relatively new and less-studied compared to other natural and social sciences, SimRAG still demonstrates promising performance in this area. This justifies the potential for adapting SimRAG to emerging domains.

5.2 Ablation Studies

Effect of Stage-I and Stage-II. Table 1 to 4 show that retrieval-oriented fine-tuning (Stage-I) significantly enhances LLM performance on QA tasks compared to the original backbone, demonstrating its effectiveness. However, further improvements become challenging after this stage. When the LLMs are fine-tuned on self-synthesized training tuples, their performance on target tasks improves even more, with an average increase of 2.21% for Llama and 3.50% for Gemma.. This suggests that, with access to a target domain corpus, LLMs can generate high-quality synthetic data, enabling self-improvement and further boosting performance.

Table 5: Results of the 5 datasets from the medical MIRAGE benchmark (Xiong et al., 2024), using DRAGON (Lin et al., 2023) as an alternative retriever.

Models	PubMedQA	BioASQ	MedQA	MedMCQA	MMLU-med	Avg.
Llama3-8B-it (2024)	57.00	81.55	55.70	55.16	65.93	63.07
SimRAG 8B	**79.60**	**91.42**	60.80	**63.88**	74.01	**73.94**
Gemma2-27B-it (2024)	58.80	89.48	57.97	55.13	76.67	67.61
SimRAG 27B	73.60	90.94	**62.29**	60.39	**79.06**	73.26

Effect of Different Retrievers. We show the performance of SimRAG using Dragon (Lin et al., 2023) as the retriever in Table 5. The results show consistent performance improvements of SimRAG over the LLM backbone, demonstrating that SimRAG is robust to different retriever choices and that its self-improvement mechanism consistently enhances performance.

5.3 Study on Pseudo-labeled Tuples

We mainly demonstrate the advantage of SimRAG in generating pseudo-labeled data from the following three perspectives.

Effect of different question generation models. To demonstrate the benefit of training on question generation and question-answering data, we compare the performance of Stage-II using different synthetic question-answer pairs. These pairs are generated either directly by Llama-3-8b-it or by an off-the-shelf QG model with T5-Large (Raffel et al., 2020) as the backbone. The results demonstrate that our approach achieves better performance on average, demonstrating the clear advantage of leveraging the fine-tuned model itself for pseudo-labeled data generation.

Effect of question filtering. We further demonstrate the advantages of question filtering in Figure 3, showing that removing low-quality data not only improves overall model performance but also accelerates the training process. It is also worth noting that even without filtering, SimRAG can achieve strong performance, suggesting that

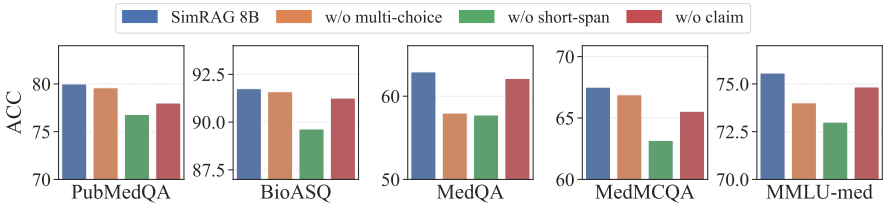

Figure 2: Effect of diverse types of generated QA pairs.

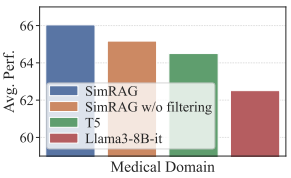

Figure 3: Effect of different generation model.

Table 6: Two case studies comparing the generated pseudo-labeled QA pairs from our fine-tuned model after Stage-I with those generated by the backbone model, Llama3-8B-it. Green and red texts represent the corresponding information between the generated QA pairs and the ground truth context.

	Case Study 1
Corpus / Type	Textbooks / Claim Verification
Context	**Title:** Histology Ross, **Text:** For its survival, the body needs to ensure a continuous delivery of energy despite highly variable supplies of nutrients from the external environment. To meet the body's energy demands when nutrient supplies are low, adipose tissue efficiently stores excess energy. The body has a limited capacity to store carbohydrate and protein, therefore energy reserves are stored within lipid droplets of adipocytes in the form of triglycerides. Triglycerides represent a dynamic form of energy storage that is added to when food intake is greater than energy expenditure and is tapped when energy expenditure is greater than food intake. The energy stored in adipocytes can be rapidly released for use at other sites in the body.
Llama3-8B-it	**Claim:** The energy stored in adipocytes can be rapidly released for use at other sites in the body. **Label:** Supported.
SimRAG (after Stage-I)	**Claim:** The body stores energy in the form of triglycerides in adipose tissue. **Label:** Supported.
	Case Study 2
Corpus / Type	Wikipedia the medical subset / Short-span QA
Context	**Title:** Cognitive behavioral therapy, **Text:** Emerging evidence for cognitive behavioral interventions aimed at reducing symptoms of depression, anxiety, and obsessive-compulsive disorder in autistic adults without intellectual disability has been identified through a systematic review. While the research was focused on adults, cognitive behavioral interventions have also been beneficial to autistic children.
Llama3-8B-it	**Question:** What age group of individuals with autism spectrum disorder, who do not have intellectual disability, have been found to benefit from cognitive behavioral interventions aimed at reducing symptoms of depression, anxiety, and obsessive-compulsive disorder, and are also a focus of? **Answer:** children.
SimRAG (after Stage-I)	**Question:** What is one of the disorders that cognitive behavioral therapy can help with? **Answer:** anxiery.

the synthetic questions generated from fine-tuned LLMs are already highly relevant to the context.

Effect of diverse question types. From Figure 2, we observe that SimRAG achieves the best performance when all three different types are included. These results justify the necessity for incorporating different task types into the fine-tuning step in Stage-II. Besides, claim verification benefits PubMedQA and BioASQ more, while multiple-choice questions boost performance on MedQA, MedMCQA, and MMLU, aligning with the question types in each dataset. Lastly, we observe that removing short-span QA leads to the largest performance drops, indicating its central role in adapting the LLM's performance towards specialized domains.

5.4 Case Studies

To better illustrate the quality of pseudo-labeled samples generated by SimRAG after Stage-I fine-tuning, we present two case studies in Table 6, comparing the samples produced by SimRAG with those from the baseline model, Llama3-8B-it.

In the first case, where the model is asked to generate a claim supported by the context, Llama3-8B-it simply selects a sentence from the context. This results in relatively simple QA pairs, making the task less challenging for Stage-II training.

In the second case, the model is tasked with generating an answer first, and then formulating a question based on the context and the answer. While Llama3-8B-it does not copy a sentence exactly, it generates a lengthy question that closely paraphrases the context. This makes the question overly dependent on the original text, making it difficult to interpret without it. Additionally, the model misinterprets the context by implying that the research was focused on children when actually adults are the focus. In contrast, after fine-tuning on answer generation and query generation in Stage-I, SimRAG generates higher-quality QA pairs that are self-contained and understandable without relying on the context. These QA pairs also present more challenging tasks, as they require deeper comprehension of the context, providing harder and more effective training data for Stage-II.

6 Conclusion

We introduce SimRAG, an instruction fine-tuning framework designed to enhance LLMs for domain-specific question-answering tasks. By equipping LLMs with joint capabilities for both question answering and question generation, SimRAG enables the generation of diverse, high-quality synthetic questions from unlabeled domain-relevant corpora. This approach facilitates effective adaptation to specialized fields, where distribution shifts and limited domain-specific data typically pose challenges. Extensive experiments across 11 datasets in three domains show that SimRAG consistently outperforms baseline models, demonstrating its effectiveness in tackling the challenges of retrieval-augmented, domain-specific question-answering tasks.

Limitation

While SimRAG demonstrates notable improvements, there are some limitations to our approach:
Single Round Pseudo-Label Generation: Our current method relies on a single round of query generation from the corpus, which may restrict the refinement of pseudo label quality. Iterative refinement of generated synthetic queries could potentially lead to better results.
Additional Training Time: The incorporation of synthetic query generation and filtering adds time complexity compared to baseline models, which may affect efficiency in environments with limited computational resources. However, we would like to note that our method *will not increase the inference time complexity* compared to the existing RAG approaches with the same backbone models.
Stronger Query Generation Models: Although we achieved strong performance with Llama3 8B and Gemma2 27B models, leveraging more powerful query generation models, such as Llama-3.1-70B-it (Meta-AI, 2024), could yield further gains. However, using larger models would incur higher computational costs beyond our current budget.

Acknowledgement

This research was partially supported by the Emory Global Diabetes Center of the Woodruff Sciences Center, Emory University.

References

Asma Ben Abacha, Eugene Agichtein, Yuval Pinter, and Dina Demner-Fushman. 2017. Overview of the medical question answering task at trec 2017 liveqa. In *TREC*, pages 1–12.

Asma Ben Abacha, Yassine Mrabet, Mark Sharp, Travis R Goodwin, Sonya E Shooshan, and Dina Demner-Fushman. 2019. Bridging the gap between consumers medication questions and trusted answers. In *MEDINFO 2019: Health and Wellbeing e-Networks for All*, pages 25–29. IOS Press.

Anthropic. 2023. Model card and evaluations for claude models.

Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. 2020. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International joint conference on neural networks (IJCNN)*, pages 1–8. IEEE.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *ICLR*.

Max Bartolo, Tristan Thrush, Robin Jia, Sebastian Riedel, Pontus Stenetorp, and Douwe Kiela. 2021. Improving question answering model robustness with synthetic adversarial data generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8830–8848, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA. Association for Computational Linguistics.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.

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

Daixuan Cheng, Yuxian Gu, Shaohan Huang, Junyu Bi, Minlie Huang, and Furu Wei. 2024. Instruction pre-training: Language models are supervised multitask learners. *arXiv preprint arXiv:2406.14491*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. 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 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv preprint*, abs/1803.05457.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free Dolly: Introducing the world's first truly open instruction-tuned llm.

Pradeep Dasigi, Nelson F. Liu, Ana Marasović, Noah A. Smith, and Matt Gardner. 2019. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5925–5932, Hong Kong, China. Association for Computational Linguistics.

Guanting Dong, Yutao Zhu, Chenghao Zhang, Zechen Wang, Zhicheng Dou, and Ji-Rong Wen. 2024. Understand what llm needs: Dual preference alignment for retrieval-augmented generation. *arXiv preprint arXiv:2406.18676*.

Jingfei Du, Edouard Grave, Beliz Gunel, Vishrav Chaudhary, Onur Celebi, Michael Auli, Veselin Stoyanov, and Alexis Conneau. 2021. Self-training improves pre-training for natural language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5408–5418, Online. Association for Computational Linguistics.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. 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 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

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*, abs/2312.10997.

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.

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, et al. 2023. Reinforced self-training (rest) for language modeling. *ArXiv preprint*, abs/2308.08998.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. Hipporag: Neurobiologically inspired long-term memory for large language models. *arXiv preprint arXiv:2405.14831*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2022. Unnatural instructions: Tuning language models with (almost) no human labor. *ArXiv preprint*, abs/2212.09689.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. 2024. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7036–7050, Mexico City, Mexico. Association for Computational Linguistics.

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.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

John Snow Labs. 2024. Jsl-medllama-3-8b-v2.0. https://huggingface.co/johnsnowlabs/JSL-MedLlama-3-8B-v2.0.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, et al. 2023. Soda: Million-scale dialogue distillation with social commonsense contextualization. In *EMNLP*.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Minh Nguyen, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Alexandrovich Glushkov, Arnav Varma Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Julian Mattick. 2023. Openassistant conversations - democratizing large language model alignment. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *ArXiv preprint*, abs/2402.10373.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Omer Levy, Luke Zettlemoyer, Jason E Weston, and Mike Lewis. 2024. Self-alignment with instruction backtranslation. In *The Twelfth International Conference on Learning Representations*.

Kevin Lin, Oyvind Tafjord, Peter Clark, and Matt Gardner. 2019. Reasoning over paragraph effects in situations. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 58–62, Hong Kong, China. Association for Computational Linguistics.

Sheng-Chieh Lin, Akari Asai, Minghan Li, Barlas Oguz, Jimmy Lin, Yashar Mehdad, Wen-tau Yih, and Xilun Chen. 2023. How to train your dragon: Diverse augmentation towards generalizable dense retrieval. In *Findings of EMNLP*.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Richard James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. 2024. RA-DIT: Retrieval-augmented dual instruction tuning. In *ICLR*.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3622–3628. ijcai.org.

Linqing Liu, Patrick Lewis, Sebastian Riedel, and Pontus Stenetorp. 2022. Challenges in generalization in open domain question answering. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2014–2029, Seattle, United States. Association for Computational Linguistics.

Xiao Liu, Hanyu Lai, Hao Yu, Yifan Xu, Aohan Zeng, Zhengxiao Du, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. Webglm: Towards an efficient web-enhanced question answering system with human preferences. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4549–4560.

Zihan Liu, Wei Ping, Rajarshi Roy, Peng Xu, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Chatqa: Surpassing gpt-4 on conversational qa and rag. In *NeurIPS*.

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

Meta-AI. 2024. Llama 3 model card.

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

John Miller, Karl Krauth, Benjamin Recht, and Ludwig Schmidt. 2020. The effect of natural distribution shift on question answering models. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 6905–6916. PMLR.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*.

OpenAI. 2022. Introducing ChatGPT.

OpenAI. 2023. GPT-4.

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. 2023. Fine-tuning or retrieval? comparing knowledge injection in llms. *ArXiv preprint*, abs/2312.05934.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *CHIL*.

Richard Yuanzhe Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason Weston. 2024. Iterative reasoning preference optimization. In *NeurIPS*.

Jungsoo Park, Sewon Min, Jaewoo Kang, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. FaVIQ: FAct verification from information-seeking questions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5154–5166, Dublin, Ireland. 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. *Journal of machine learning research*, 21(140):1–67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Chuck Rosenberg, Martial Hebert, and Henry Schneiderman. 2005. Semi-supervised self-training of object detection models. In *2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05)*.

Tobias Schimanski, Jingwei Ni, Mathias Kraus, Elliott Ash, and Markus Leippold. 2024. Towards faithful and robust LLM specialists for evidence-based question-answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1913–1931, Bangkok, Thailand. Association for Computational Linguistics.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024a. Replug: Retrieval-augmented black-box language models. In *NAACL*.

Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Hang Wu, Carl Yang, and May D Wang. 2024b. Medadapter: Efficient test-time adaptation of large language models towards medical reasoning. In *EMNLP*.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiaoshuai Song, Muxi Diao, Guanting Dong, Zhengyang Wang, Yujia Fu, Runqi Qiao, Zhexu Wang, Dayuan Fu, Huangxuan Wu, Bin Liang, et al. 2024. Cs-bench: A comprehensive benchmark for large language models towards computer science mastery. *ArXiv preprint*, abs/2406.08587.

Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. 2024. DRAGIN: Dynamic retrieval augmented generation based on the real-time information needs of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12991–13013, Bangkok, Thailand. Association for Computational Linguistics.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *ArXiv preprint*, abs/2408.00118.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*.

Tu Vu, Minh-Thang Luong, Quoc Le, Grady Simon, and Mohit Iyyer. 2021. STraTA: Self-training with task augmentation for better few-shot learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5715–5731, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

David Wadden, Kejian Shi, Jacob Morrison, Aakanksha Naik, Shruti Singh, Nitzan Barzilay, Kyle Lo, Tom Hope, Luca Soldaini, Shannon Zejiang Shen, et al. 2024. Sciriff: A resource to enhance language model instruction-following over scientific literature. *ArXiv preprint*, abs/2406.07835.

Haoyu Wang, Tuo Zhao, and Jing Gao. 2024. Blendfilter: Advancing retrieval-augmented large language models via query generation blending and knowledge filtering. In *EMNLP*.

Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych. 2022. GPL: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2345–2360, Seattle, United States. Association for Computational Linguistics.

Yaqing Wang, Subhabrata Mukherjee, Haoda Chu, Yuancheng Tu, Ming Wu, Jing Gao, and Ahmed Hassan Awadallah. 2021. Meta self-training for few-shot neural sequence labeling. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1737–1747.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023a. Self-instruct: Aligning language models with self-generated instructions. In *ACL*.

Yubo Wang, Xueguang Ma, and Wenhu Chen. 2023b. Augmenting black-box llms with medical textbooks for clinical question answering. *ArXiv preprint*, abs/2309.02233.

Zhepei Wei, Wei-Lin Chen, and Yu Meng. 2024. Instructrag: Instructing retrieval-augmented generation with explicit denoising. *ArXiv preprint*, abs/2406.13629.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.

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

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6233–6251, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2024a. RECOMP: Improving retrieval-augmented LMs with context compression and selective augmentation. In *ICLR*.

Ran Xu, Wenqi Shi, Yue Yu, Yuchen Zhuang, Yanqiao Zhu, May D Wang, Joyce C Ho, Chao Zhang, and Carl Yang. 2024b. Bmretriever: Tuning large language models as better biomedical text retrievers. In *EMNLP*.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making retrieval-augmented language models robust to irrelevant context. In *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. In *EMNLP*.

Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Rankrag: Unifying context ranking with retrieval-augmented generation in llms. In *NeurIPS*.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. In *ICML*.

Dan Zhang, Ziniu Hu, Sining Zhoubian, Zhengxiao Du, Kaiyu Yang, Zihan Wang, Yisong Yue, Yuxiao Dong, and Jie Tang. 2024a. Sciglm: Training scientific language models with self-reflective instruction annotation and tuning. *ArXiv preprint*, abs/2401.07950.

Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, et al. 2024b. Ultramedical: Building specialized generalists in biomedicine. *arXiv preprint arXiv:2406.03949*.

Tianjun Zhang, Shishir G Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. 2024c. RAFT: Adapting language model to domain specific RAG. In *COLM*.

Xinlu Zhang, Chenxin Tian, Xianjun Yang, Lichang Chen, Zekun Li, and Linda Ruth Petzold. 2023. Alpacare: Instruction-tuned large language models for medical application. *arXiv preprint arXiv:2310.14558*.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3277–3287, Online. Association for Computational Linguistics.

A Training Data Details

We include the training dataset, the number of examples used in each stage, as well as the instruction format in Table 7. Note that in our implementation, we use the `interleave_datasets()` function to generate the final training data by merging multiple sources. The numbers presented in the table represent the sample counts of the training data from their original sources. In Stage-I, we set `stopping_strategy="all_exhausted"`, an oversampling strategy where dataset construction stops only after every sample from all datasets has been included at least once. In Stage-II, we use `stopping_strategy="first_exhausted"`, an undersampling strategy where construction stops as soon as one dataset has ran out of samples. In this case, the pseudo-labeled QA dataset is the first to run out, ensuring that all its samples are fully utilized, while only a small portion of the other datasets is included.

B Test Data Details

We evaluate on 11 datasets in total from the medical, scientific and computer science domain. (1) **Medical:**

- MMLU-med (Hendrycks et al., 2021) is a subset of six tasks related to biomedicine, including anatomy, clinical knowledge, professional medicine, human genetics, college medicine, and college biology. It contains 1089 questions in total.

- MedMCQA (Pal et al., 2022) includes multiple-choice questions derived from Indian medical entrance exams, covering 2400 healthcare topics across 21 medical subjects. We use the 4,183-question development set from MedMCQA, as the test set lacks provided ground truths.

- MedQA (Jin et al., 2021) is collected from the US Medical Licensing Examination, containing 1273 four-option multiple-choice questions focused on real-world scenarios from professional medical board exams.

- BioASQ (Tsatsaronis et al., 2015) includes 618 questions constructed from biomedical literature without providing the ground truth snippets, challenging RAG systems to infer answers independently.

Table 7: The blending ratio of different datasets with their specific prompt format in Stage-I and Stage-II fine-tuning. For Stage-II Pseudo-labeled QA Samples, the two numbers represent the # sample for the Llama and Gemma backbones, respectively.

Dataset	Specific Instruction	Stage-I # Samples	Stage-I Blending Ratio	Stage-II # Samples	Stage-II Blending Ratio
Instruction Fine-tuning					
ChatQA SFT Data	—	60000	0.18	128000	0.12
Question Answering					
DROP	Answer the following question with a short span.	12000	0.034	29195	0.04
NarrativeQA		12000	0.034	40000	0.04
Quoref		4800	0.014	10996	0.015
ROPES		4800	0.014	10924	0.015
Squad1.1		16000	0.045	40000	0.035
Squad2.0		16000	0.045	52474	0.05
OpenbookQA	Answer the following question by selecting one of the provided options with A, B, C, or D. Please answer with the capitalized alphabet only, without adding any extra phrase or period.	2000	0.006	82092	0.005
LogiQA		4000	0.012	7376	0.006
NQ	Answer the following question with a short phrase.	16000	0.045	46426	0.04
TatQA-arithmetic	Answer the following question with a number from context or the math arithmetic using +,-,*, or /.	8325	0.045	24975	0.034
TatQA-others	Answer the following question with a short span, or a full and complete answer.	3176	0.023	9528	0.013
WebGLM	Please give a full and complete answer for the question using only the provided search results (some of which might be irrelevant) and cite them properly. Use an unbiased and journalistic tone. When citing several search results, use [1][2][3].	12000	0.034	43579	0.023
StrategyQA	Answer the following question with Yes or No.	1526	0.005	4578	0.006
BoolQ		4000	0.012	9427	0.013
FaVIQ	Answer the following question with Yes or No. Is the statement {claim} correct?	2000	0.006	10906	0.01
FEVER		2000	0.006	10444	0.01
Pseudo-labeled Question Answering					
Short-span QA	Answer the following question with a short span.	—	—	150,000 / 45,000	0.2625
Multiple-choice QA	Answer the following question by selecting one of the provided options with A, B, C, or D. Please answer with the capitalized alphabet only, without adding any extra phrase or period.	—	—	50,000 / 15,000	0.0875
Claim Verification	Answer the following question with Yes or No. Is the statement {claim} correct?	—	—	100,000 / 30,000	0.175
Answer Generaion					
Squad1.1	Based on the context, generate candidate spans within the passage that are likely to be answers to a question. Separate different candidate answers with a semicolon (';').	18877	0.063	—	—
Squad2.0		18863	0.059	—	—
DROP		4984	0.023	—	—
WebQuestions		1084	0.012	—	—
Query Generaion					
NQ	Based on the context, please generate a question. The answer to the question should be {answer}.	20000	0.068	—	—
Squad1.1		20000	0.068	—	—
StrategyQA		131	0.023	—	—
WebQuestions		24000	0.068	—	—
FaVIQ	Based on the context, please generate a claim that can be supported/refuted by the context.	10000	0.028	—	—
FEVER		10000	0.028	—	—

- PubMedQA (Jin et al., 2019) is a biomedical research QA dataset consisting of 1000 manually annotated questions based on PubMed abstracts. Answers in PubMedQA are structured as yes/no/-maybe to reflect the validity of the questions.

- LiveQA (Abacha et al., 2017) and MedicationQA (Abacha et al., 2019) are two QA datasets focusing on answering consumer health questions about medications, including 100 and 674 question-answer pairs, respectively.

(2) **Scientific:**

- SciQ (Welbl et al., 2017) is a scientific question-answering dataset containing 13,679 crowd-sourced science exam questions about Physics, Chemistry, and Biology, among others.

- ARC-easy/challenge (Clark et al., 2018) contains 7,787 authentic multiple-choice science questions at the grade-school level, designed to foster advanced question-answering research. The dataset is divided into a Challenge Set, with questions that stumped both a retrieval-based and a word co-occurrence algorithm, and an Easy Set.

- MMLU-Sci (Hendrycks et al., 2021) is the Massive Multitask Language Understanding dataset, designed to test a wide range of language understanding abilities across 57 tasks. In this work, we select 14 subjects to ensure the evaluation is not limited to certain fields.

(3) **Computer Science:**

- CS-Bench (Song et al., 2024) is a recently-proposed benchmark specifically designed to assess the performance of large language models (LLMs) in computer science. It contains around 5,000 carefully selected test samples that span 26 subfields within four major areas of computer science, covering various task forms and divisions of knowledge and reasoning.

C Baseline Descriptions

- Self-RAG (Asai et al., 2024) utilizes instruction fine-tuning to adaptively retrieve passages based on the question and determine if the passage contains useful information for answering the question.

- ChatQA (Liu et al., 2024) is a fine-tuning pipeline tailored for RAG and conversational QA tasks via aggregating multiple QA and dialogue datasets.

- RAFT (Zhang et al., 2024c) is a domain-specific fine-tuning approach that incorporates top-k passages as context during fine-tuning, helping to address discrepancies between training and testing data.

- EvidenceRAG (Schimanski et al., 2024) leverage off-the-shelf LLMs (GPT-4) to generate context-aware question answering datasets, which is then used to fine-tune the student model.

D Additional Experimental Results

We list the per-task results of MMLU-sci in Table 8.

E Prompt Details

E.1 Answer Generation

```
[System]

[Context]

Based on the context, generate
several candidate spans within
the passage that are likely to be
 answers to a question. The
answers can be entities, verbs or
 even numbers. Make sure that the
 answers are different and
diverse. Separate different
candidate answers with a
semicolon (';').
```

E.2 Query Generation

```
[System]

[Context]

Based on the context, please
generate a question that is
relevant to the information
provided. The question should
stand alone and not refer back to
 the context explicitly. The
question should be clear and
understandable without needing
the context. The answer to the
question should be [Answer].
```

E.3 Inference

```
[System]

[Top 10 Contexts]

[Specific Instruction]

[Question]
```

The [Specifc Instruction] for each evaluation dataset depends on their question type and can refer to those in Table 7.

Table 8: Results of our proposed method and baselines in the scientific domain.

Models	**astronomy**	**college biology**	**college chemistry**	**college physics**	**computer security**	**high school geography**	**high school macroeconomics**	**high school microeconomics**	**high school psychology**	**high school US history**	**high school world history**	**human sexuality**	**nutrition**	**virology**	**Avg.**
Metrics	ACC	ACC	ACC	ACC	ACC	ACC	ACC	ACC	ACC	ACC	ACC	ACC	ACC	ACC	—
Proprietary LLMs, For Reference Only															
GPT-3.5 (OpenAI, 2022)	66.45	65.28	35.00	46.53	65.00	77.27	91.54	64.29	83.12	78.43	72.15	70.99	66.01	47.59	66.40
GPT-4 (OpenAI, 2023)	93.42	93.75	61.00	73.27	91.00	94.95	97.95	94.54	96.15	95.59	94.51	93.13	89.22	56.02	87.46
Scientific LLMs															
SciTulu 7B (Wadden et al., 2024)	69.74	63.89	31.00	18.63	62.00	70.20	56.58	57.08	77.43	53.06	57.38	65.65	54.90	45.78	55.95
SciTulu 70B (Wadden et al., 2024)	83.55	80.56	36.00	28.43	83.00	89.39	80.26	79.83	91.19	77.55	77.22	78.63	68.95	50.60	71.80
Retrieval-Augmented LLMs															
Self-RAG 13B (Asai et al., 2024)	55.26	58.33	24.00	21.57	60.00	61.11	32.89	45.49	67.89	58.67	58.23	53.44	43.79	40.96	48.69
ChatQA 8B (Liu et al., 2024)	60.53	54.17	29.00	33.33	70.00	64.65	51.32	58.37	74.86	49.49	54.85	59.54	57.19	45.18	54.46
ChatQA 70B (Liu et al., 2024)	82.89	79.17	46.00	48.04	83.00	84.85	80.26	84.98	91.74	**86.73**	82.28	74.05	77.78	51.20	75.21
Backbone: Llama3-8B-Instruct															
Llama3-8B-it (Meta-AI, 2024)	78.29	71.53	38.00	40.20	83.00	82.32	63.16	72.96	84.04	65.31	72.15	69.47	70.26	49.40	67.15
RAFT 8B (Zhang et al., 2024c)	80.26	75.69	37.00	42.16	84.00	79.80	65.79	74.68	83.67	72.45	77.22	73.28	71.24	51.81	69.22
EvidenceRAG 8B (Schimanski et al., 2024)	77.63	78.47	44.00	45.10	85.00	84.85	72.37	74.68	86.24	74.49	79.32	74.05	74.84	51.20	71.59
SimRAG 8B	85.53	81.94	47.00	50.98	**88.00**	89.90	76.32	84.55	92.66	83.16	81.43	84.73	**81.37**	54.82	77.31
w/o Stage II	84.87	81.25	**49.00**	49.02	87.00	88.89	73.68	82.83	90.64	80.61	81.01	83.21	79.41	51.81	75.95
Backbone: Gemma2-27B-Instruct															
Gemma2-27B-it (Team et al., 2024)	82.89	84.03	47.00	55.88	84.00	89.39	77.63	81.12	91.93	80.61	84.81	81.68	72.22	52.40	76.11
RAFT 27B (Zhang et al., 2024c)	84.87	88.89	47.00	63.73	86.00	90.91	**86.84**	84.55	93.58	81.12	85.65	81.68	76.47	51.81	78.79
EvidenceRAG 27B (Schimanski et al., 2024)	84.87	87.50	**49.00**	60.78	86.00	91.41	**86.84**	85.41	93.94	81.63	86.08	81.68	76.80	51.81	78.84
SimRAG 27B	**90.13**	**91.67**	**49.00**	**68.63**	87.00	**92.42**	85.53	**87.98**	**95.05**	84.18	**86.92**	**85.50**	78.43	**55.42**	**81.28**
w/o Stage II	84.21	87.50	**49.00**	59.80	84.00	89.90	84.21	82.83	93.58	83.16	86.50	81.68	76.14	54.82	78.38