Select to Know: An Internal-External Knowledge Self-Selection Framework for Domain-Specific Question Answering

Anonymous ACL submission

Abstract

Large Language Models (LLMs) perform well in general QA but often struggle in domain-specific scenarios. Retrieval-Augmented Generation (RAG) introduces external knowledge but suffers from hallucinations and latency due to noisy retrievals. Continued pretraining internalizes domain knowledge but is costly and lacks cross-domain flexibility. We attribute this challenge to the long-tail distribution of domain knowledge, which causes partially internalized yet useful knowledge to be underutilized. We further argue that knowledge acquisition should be progressive, mirroring human learning: first understanding concepts, then applying them to complex reasoning. To address this, we propose *Select2Know* (S2K), a cost-effective framework that internalizes domain knowledge through an internal-external knowledge self-selection strategy and selective supervised fine-tuning. We also introduce a structured reasoning data generation pipeline and integrate GRPO to enhance reasoning ability. Experiments on medical, law, and financial QA benchmarks show that S2K consistently outperforms existing methods and matches domain-pretrained LLMs with significantly lower cost.

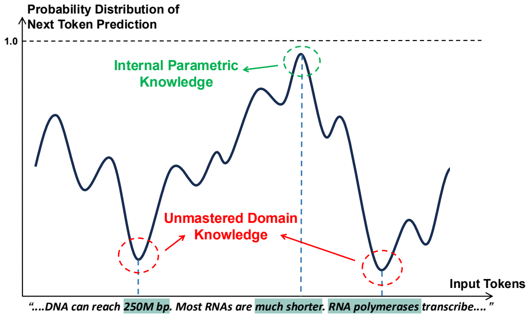


Figure 1: Visualization of token-level prediction probabilities. Low-probability tokens indicate unmastered domain knowledge, while high-probability tokens reflect internal parametric knowledge. This highlights the need for integrating internal and external knowledge in domain adaptation. (Note: Schematic illustration; see Appendix A.1 for real examples.)

1 Introduction

With the rapid advancement of large language models (LLMs), their effectiveness in general question answering has been widely validated (Devlin et al., 2019; Brown et al., 2020; Lewis et al., 2020; Shailendra et al., 2024). However, LLMs still exhibit noticeable performance gaps in domain-specific QA tasks (Yang et al., 2023; Yue, 2025). To address these challenges, a variety of approaches have been explored to improve domain-specific QA (DSQA) performance.

A common solution is the use of Retrieval-Augmented Generation (RAG) (Lewis et al., 2020; Press et al., 2023; Asai et al., 2023; He et al., 2024), where an retriever is used to access external knowledge from a domain corpus. While RAG helps incorporate up-to-date information, it introduces extra latency and computation due to redundant retrievals. Additionally, distribution mismatches may lead the retriever to return irrelevant or conflicting information, increasing the risk of hallucinations (Rawte et al., 2023; Ji et al., 2023; Ye et al., 2023; Maynez et al., 2020; Xu et al., 2024).

Another line of research focuses on enhancing domain adaptation through continued pretraining (Labrak et al., 2024; Qiu et al., 2024; Shu et al., 2024; Li et al., 2024; Chen et al., 2023). These methods can achieve strong performance, but they are extremely resource-intensive and often lack transferability to other domains. (BioMistral (Labrak et al., 2024) requires training on a corpus of three billion tokens.)

We argue that the fundamental reason behind LLMs' poor performance in DSQA lies in the long-tail distribution of domain knowledge in pretraining data. As illustrated in Figure 1, LLMs have already internalized parts of domain knowledge during pretraining. While this knowledge is often incomplete, it can complement or even correct external domain inputs, making external-only methods suboptimal. Furthermore, we believe knowledge acquisition should follow a human-inspired staged progression—first achieving conceptual comprehension, then advancing to complex reasoning.

Building on this insight, we propose a low-cost post-training framework, ***S**elect**2K**now* (S2K), for domain-specific question answering, which integrates both internal parametric knowledge and external domain knowledge. Specifically, we first introduce a token-level internal-external *knowledge self-selection* strategy to construct fusion training data. We then propose *Selective Supervised Fine-Tuning (Selective SFT)* to guide the model toward focusing on domain knowledge it has not yet mastered. In addition, we design a *structured data generation pipeline* to efficiently produce high-quality reasoning data, and incorporate *Group Relative Policy Optimization (GRPO)* (Shao et al., 2024) to enhance the model's ability to apply learned knowledge to real-world reasoning tasks. Our main contributions are as follows:

- We propose a token-level knowledge self-selection strategy to fuse internal parametric knowledge and external domain knowledge.
- We propose a low-cost post-training framework to boost LLM performance on DSQA.
- Experiments across the medicine, law, and finance demonstrate that S2K matches pre-trained LLMs with significantly lower training cost.

2 Problem Definition

We aim to design a general pipeline that enables LLMs to efficiently generalize to domain-specific QA tasks with minimal cost. To closely reflect real-world scenarios, we make the following assumptions: (1) No existing QA training datasets are available in the target domain. (2) The only accessible resource is a collection of unstructured domain-specific corpus $\mathcal{D} = \{d_1, d_2, ..., d_n\}$, such as news, textbooks, regulatory documents, etc. (3) A pre-trained general LLM $\mathcal{M}_0$ (e.g., LLaMA(Touvron et al., 2023; Grattafiori et al., 2024), Qwen(Yang et al., 2024a,b)) is used as the foundation.

Our goal is to develop a pipeline $\mathcal{P}$ such that the resulting domain-adapted model $\mathcal{M}_\mathcal{D} = \mathcal{P}(\mathcal{M}_0, \mathcal{D})$ achieves strong performance on the domain QA task $\mathcal{T}_{\mathcal{QA}}$. Formally, we aim for $\mathrm{Perf}(\mathcal{M}_\mathcal{D}, \mathcal{T}_{\mathcal{QA}}) \gg \mathrm{Perf}(\mathcal{M}_0, \mathcal{T}_{\mathcal{QA}})$, where $\mathrm{Perf}(\cdot)$ denotes the evaluation performance on domain QA tasks.

3 Methods

We introduce S2K, a low-cost post-training framework for adapting general LLMs to domain-specific QA. As illustrated in Figure 2, S2K first extracts question-style meta knowledge from raw domain corpora (Section 3.1.1). We then design a token-level self-selection mechanism to fuse internal and external knowledge (Section 3.1.2), complemented by Selective SFT, which guides the model to focus on unfamiliar domain knowledge (Section 3.2). We further introduce structured reasoning data generation pipeline (Section 3.1.3), and incorporate GRPO to enhance the model's reasoning ability for complex real-world scenarios (Section 3.3).

3.1 Domain Knowledge Generation

3.1.1 Meta Knowledge

As described in Section 2, we construct domain QA data by first extracting question-style meta knowledge from raw domain corpora $\mathcal{D}$. Since such corpora are often redundant and unstructured, containing irrelevant details such as timestamps or publisher metadata, we first cleaning the data to remove non-informative content, then segment the corpus into token-balanced chunks using NLTK (Bird, 2006) to preserve semantic coherence. For each chunk $d_i \in \mathcal{D}$, we prompt a LLM (e.g., DeepSeek-v3 (Liu et al., 2024) or GPT-4o (Hurst et al., 2024)) to generate a knowledge question. Formally, the question-style meta knowledge is defined as:

$$\mathcal{Q}_i = f_{\text{prompt}}(\mathcal{L}, d_i) \tag{1}$$

where $\mathcal{L}$ denotes the LLM used for prompting, f_{prompt} is the prompting process, and $\mathcal{Q}_i$ is the meta question. Detailed prompts are provided in Appendix A.5.

3.1.2 Internal-External Fusion Knowledge

An intuitive approach to domain knowledge training is using answers generated from question-style meta knowledge and their corresponding text chunks. However, these answers rely only on external documents, which may introduce noise and ignore the model's internal knowledge. To address this, we propose a token-level internal-external **knowledge self-selection** strategy. Specifically, we make internal and external knowledge explicit

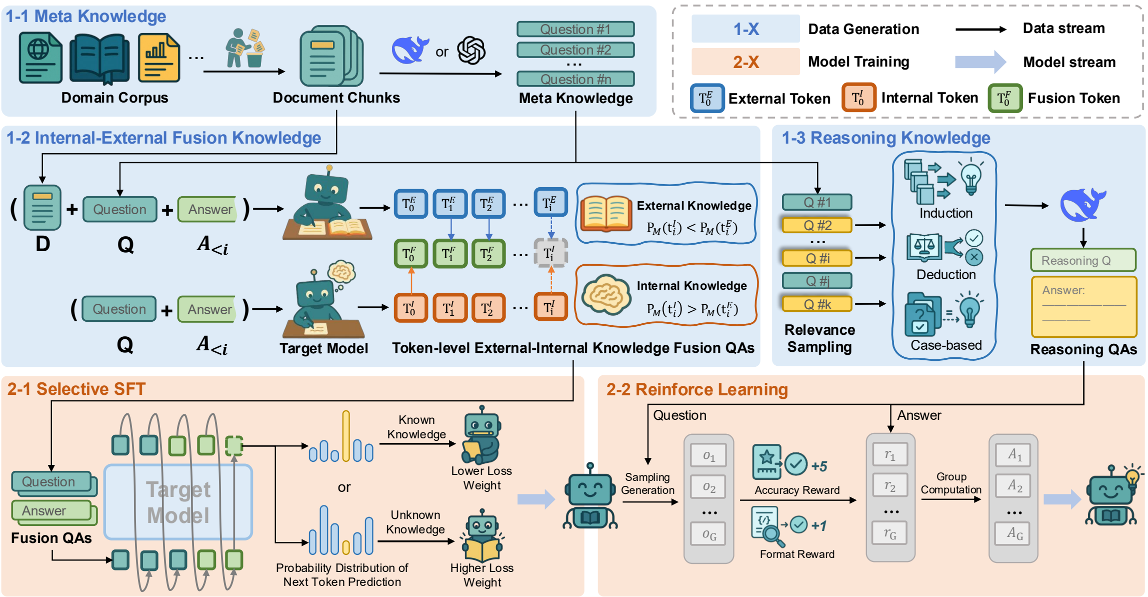


Figure 2: Overview of S2K, a low-cost post-training framework for domain-specific QA. The method comprises: data generation (1-X) and model training (2-X). In data generation, question-style meta knowledge is extracted from domain corpora, followed by token-level fusion of internal and external knowledge, and reasoning QA construction via relevance-based sampling and structured prompts. In model training, Selective SFT emphasizes unmastered knowledge using token-level uncertainty, while GRPO-based reinforcement learning enhances reasoning.

through two parallel inference settings: one with both the question and its supporting text chunk ($Q + D$) as context, representing external knowledge ($A^E = P_M(Q, D)$), and one with the question alone (Q) as context, reflecting internal knowledge ($A^I = P_M(Q)$). Here, M denotes the target model, and $P_M(\cdot)$ represents its inference process.

The key challenge is determining how to fuse A^E and A^I at the token level. We propose a simple yet effective strategy based on the target model's predicted probabilities: without loss of generality, for token t_i, if the model assigns a higher probability to it under the internal setting than under the external one, we select the internal token; otherwise, the external token. Formally:

$$t_i^F = \begin{cases} t_i^I, & \text{if} P_M(t_i^I \mid Q, A_{<i}^F) > P_M(t_i^E \mid Q, D, A_{<i}^F) \\ t_i^E, & \text{otherwise} \end{cases} \tag{2}$$

Here, $A_{<i}^F = \{t_0^F, t_1^F, t_2^F, \ldots, t_{i-1}^F\}$, which ensures two key properties: (1) the final answer fused from internal and external knowledge remains coherent and readable, and (2) the only difference between the two inference settings is whether the external document D is provided.

In practice, selecting knowledge token by token can be overly greedy and lead to locally optimal answers. To address this, we adopt a window-based generation strategy, model generates multiple tokens (W) per step and selects between internal and external knowledge based on their average log-probabilities within the window. Meanwhile, to further mitigate overconfidence, we apply a scaling factor C to favor external knowledge when appropriate. Moreover, we use log-probabilities instead of raw probabilities to enhance comparability across tokens. The final implementation is formalized as:

$$t_{i:i+W}^F = \begin{cases} t_{i:i+W}^I, & \text{if } \frac{1}{W} \sum_{j=0}^{W-1} \log P_M(t_{i+j}^I \mid Q, A_{<i}^F) \geq \\ & \quad \frac{1}{W} \sum_{j=0}^{W-1} \log P_M(t_{i+j}^E \mid Q, D, A_{<i}^F) \\ & \quad + C \\ t_{i:i+W}^E, & \text{otherwise} \end{cases} \tag{3}$$

3.1.3 Reasoning Knowledge

Real-world domain scenarios often require reasoning across multiple knowledge points. To simulate this, we adopt a **relevance-based sampling** strategy: for each question and its corresponding document chunk, we retrieve the top 10 related question-chunk pairs, which serve as the basis for constructing complex reasoning queries.

To ensure the diversity and quality of the reasoning data, we propose a **structured data generation pipeline** that classifies reasoning types into three categories: (1) **Deductive** Reasoning follows a top-down logical process, applying general knowledge points to specific reasoning cases, (2) **Inductive** Reasoning works in the opposite direction, deriving general patterns or principles from multiple specific instances, (3) **Case-based** Reasoning involves analogical thinking, where the solution to a new problem is inferred by comparing it with previously encountered similar cases. For each type, we design tailored prompts to guide the LLM in combining the sampled questions with relevant document chunks to form coherent, multi-step reasoning QA pairs. This structured approach enables controlled and diverse QA synthesis, enhancing logical depth while providing a general pipeline for efficiently generating high-quality reasoning data. Details and examples for each reasoning type are provided in Appendix A.4 and A.5. The overall data generation process is illustrated in Algorithm 1.

3.2 Internal–External Knowledge Fusion Training

In the internal-external fusion data (Section 3.1.2), part of the knowledge is already embedded in the internal parameters of the model. Therefore, applying standard supervised fine-tuning can lead to inefficient training and slower adaptation to new knowledge. To mitigate this, we propose **Selective Supervised Fine-Tuning (Selective SFT)**, which leverages the model's token-level uncertainty. Tokens with higher uncertainty, indicating unfamiliar or novel knowledge, are given greater weight during optimization, while confident predictions contribute less to the loss.

To quantify the model's uncertainty, we compute the per-token entropy based on output logits. The entropy H_t for each token is defined as:

$$H_t = -\sum_{v=1}^{V} p_t(v) \log p_t(v) \tag{4}$$

where $p_t(v)$ is the predicted probability of token v at step t, and V is the vocabulary size. To allow comparison across models or vocabularies, we normalize H_t by the maximum entropy $\log V$.

The token-wise weight factor ω_t is defined as:

$$\omega_t = (1 - \text{correct}_t) + \text{correct}_t \cdot \frac{H_t}{\log V} \tag{5}$$

where correct_t is an indicator function that equals 1 if the token prediction is correct, and 0 otherwise.

Algorithm 1 Domain Knowledge Generation

Input: Domain corpus $\mathcal{D}$, LLM M, Retriever R, Max answer length L, Window size W, Margin C, Reasoning types $\mathcal{R}_t$

```
// Step 1: Meta Knowledge Extraction
Clean and segment $\mathcal{D}$ into token-balanced chunks $\{d_i\}$
for each chunk $d_i$ do
    Generate meta questions $\{q_i\}$ from $d_i$
end for
// Step 2: Internal-External Fusion Knowledge
for each question $q$ and chunk $d$ do
    Init Context$_E \leftarrow (q, d)$, Context$_I \leftarrow (q)$, $G \leftarrow \emptyset$
    while $|G| < L$ do
        Generate $T_E$, $T_I$ under Context$_E$, Context$_I$
        Compute avg. log-probs $p_E$, $p_I$
        Select $T_I$ if $p_I \geq p_E + C$, else select $T_E$
        update Context$_E$, Context$_I$
        if EOS token in $G$ then break
        end if
    end while
end for
// Step 3: Reasoning Knowledge
for each question $q$ in meta knowledge set do
    Retrieve $k$ relevant pairs $\{(q_i, d_i)\}_{i=1}^{k}$ by $R$
    for each reasoning type $r$ in $\mathcal{R}_t$ do
        Construct prompt $\mathcal{P}_r$ according to type $r$
        Generate QA pair $(q', a')$ using $\mathcal{P}_r$, $\{(q_i, d_i)\}_{i=1}^{k}$
    end for
end for
```

Output: Internal-external Fusion QAs and Reasoning QAs

The final loss is computed as a weighted negative log-likelihood (NLL):

$$\mathcal{L} = \frac{1}{N} \sum_{t=1}^{T} \omega_t \cdot \text{NLL}_t \tag{6}$$

where N is the number of valid tokens and NLL_t denotes the negative log-likelihood at step t. This uncertainty-aware objective prioritizes unmastered external knowledge and avoids redundant updates, enabling more efficient fine-tuning.

3.3 Reasoning-Enhanced Training

After acquiring domain knowledge, we apply GRPO, a critic-free reinforcement learning method, to improve the reasoning capabilities of the LLM. Following raw GRPO, we design an accuracy reward and a format reward. The accuracy reward (R_{acc}) has two cases: +5 for a fully correct answer and 0 for an incorrect one. The format reward (R_{fmt}) includes three cases: +1 for strictly following the "<think>...</think>...ANSWER" format, 0 for a general formatting error, and –0.5 if "ANSWER" is generated multiple times, which indicating a potential reward-hacking behavior, where the model outputs multiple candidate answers to maximize reward. The final reward is the sum of both: $R = R_{\text{acc}} + R_{\text{fmt}}$.

Method		MedQA			JECQA			FinanceIQ		
		Avg@5	**Cons@5**	**Pass@5**	**Avg@5**	**Cons@5**	**Pass@5**	**Avg@5**	**Cons@5**	**Pass@5**
Zero-Shot		33.5	38.3	67.6	15.9	18.0	39.5	18.0	17.7	62.2
Few-Shot	1-shot	33.6$_{+0.1}$	36.2$_{-2.1}$	68.1$_{+0.5}$	15.2$_{-0.7}$	16.7$_{-1.3}$	39.7$_{+0.2}$	17.5$_{-0.5}$	16.6$_{-1.1}$	60.9$_{-1.3}$
	3-shot	33.0$_{-0.5}$	35.7$_{-2.6}$	67.6$_{+0.0}$	12.3$_{-3.6}$	11.2$_{-6.8}$	34.9$_{-4.6}$	16.2$_{-1.8}$	14.0$_{-3.7}$	58.1$_{-4.1}$
	5-shot	33.8$_{+0.3}$	36.3$_{-2.0}$	67.1$_{-0.5}$	13.8$_{-2.1}$	13.2$_{-4.8}$	37.9$_{-1.6}$	16.0$_{-2.0}$	14.4$_{-3.3}$	57.3$_{-4.9}$
RAG	Naive	34.2$_{+0.7}$	38.3$_{+0.0}$	65.9$_{-1.7}$	6.1$_{-9.8}$	4.7$_{-13.3}$	17.6$_{-21.9}$	11.8$_{-6.2}$	5.4$_{-12.3}$	46.6$_{-15.6}$
	Self-Ask	20.3$_{-13.2}$	21.7$_{-16.6}$	67.9$_{+0.3}$	9.4$_{-6.5}$	13.9$_{-4.1}$	18.2$_{-21.3}$	3.0$_{-15.0}$	0.3$_{-17.4}$	13.3$_{-48.9}$
	Self-RAG	23.4$_{-10.1}$	25.3$_{-13.0}$	72.7$_{+5.1}$	6.4$_{-9.5}$	14.6$_{-3.4}$	17.7$_{-21.8}$	10.1$_{-7.9}$	4.3$_{-13.4}$	41.2$_{-21.0}$
Post-Training	SFT	32.4$_{-1.1}$	35.9$_{-2.4}$	68.4$_{+0.8}$	15.3$_{-0.6}$	16.9$_{-1.1}$	42.6$_{+3.1}$	23.1$_{+5.1}$	25.1$_{+8.0}$	71.4$_{+9.2}$
	PPO	34.2$_{+0.7}$	34.8$_{-3.5}$	40.6$_{-27.0}$	18.0$_{+2.1}$	18.1$_{+0.1}$	28.6$_{-10.9}$	23.6$_{+5.6}$	25.7$_{+8.0}$	69.7$_{+7.5}$
	GRPO	36.1$_{+2.6}$	36.4$_{-1.9}$	61.4$_{-6.2}$	21.1$_{+5.2}$	21.5$_{+3.5}$	29.3$_{-10.2}$	22.6$_{+4.6}$	24.5$_{+6.8}$	72.3$_{+10.1}$
	Sel. SFT (Ours)	35.1$_{+1.6}$	39.6$_{+1.3}$	75.9$_{+8.3}$	18.6$_{+2.7}$	23.1$_{+5.1}$	42.1$_{+2.6}$	23.6$_{+5.6}$	25.5$_{+7.8}$	72.3$_{+10.1}$
	S2K (Ours)	**38.6**$_{+5.1}$	**43.4**$_{+5.1}$	**77.1**$_{+9.5}$	**26.2**$_{+10.3}$	**27.7**$_{+9.7}$	**43.6**$_{+4.1}$	**25.8**$_{+7.8}$	**27.7**$_{+10.0}$	**73.4**$_{+11.2}$

Table 1: We evaluate S2K against representative domain-specific QA enhancement methods across prompting, RAG, and post-training approaches on three benchmarks: MedQA (medicine), JECQA (law), and FinanceIQ (finance). S2K consistently outperforms other QA enhancement strategies we benchmarked, highlighting the effectiveness of internal-external knowledge fusion and two-stage training. (Sel. SFT means Selective SFT we proposed)

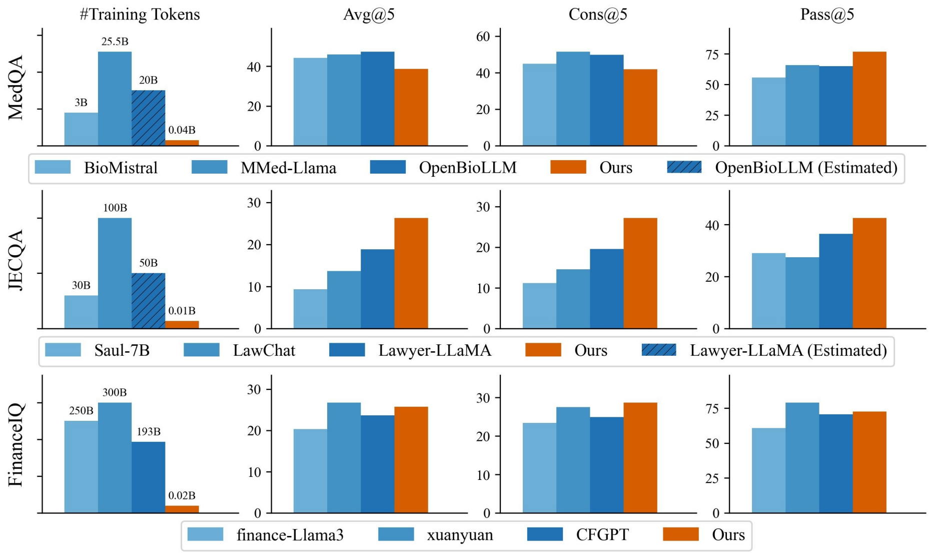


Figure 3: Compared to domain-specific LLMs pretrained on large-scale corpora, S2K reaches comparable performance using 2–3 orders of magnitude less data, demonstrating the effectiveness of internal-external knowledge fusion. Striped bars indicate estimated training tokens due to missing data from the original papers.

4 Experiments

We organize our experiments as follows: Section 4.1 details the experimental setup. Section 4.2 provides a quantitative comparison between our method and other question-answering enhancement paradigms. Section 4.3 analyzes the sensitivity of key hyperparameters, revealing underlying mechanisms of our method. Section 4.4 presents ablation study to examine the contribution of each module. Finally, Section 4.5 provides case studies illustrating the use of internal knowledge in practice.

4.1 Experiment Setup

Datasets: To evaluate the cross-domain generalization of S2K, we conduct experiments in three domains: medicine (MedQA (Jin et al., 2021)), law (JEC-QA (Zhong et al., 2020)), and finance (FinanceIQ (Zhang and Yang, 2023)). **MedQA** is a multilingual medical QA benchmark based on professional exams. Training is based on medical textbooks, and evaluation is conducted on the MedQA-USMLE subset. **JEC-QA** (Zhong et al., 2020) is a legal QA dataset derived from the Chinese National Judicial Examination. S2K is evaluated on the JEC-QA-KD subset from AGIEval (Zhong et al., 2024). **FinanceIQ** (Zhang and Yang, 2023) is a Chinese financial QA dataset with multiple-choice questions across diverse topics. Training data is sampled from corresponding FinCorpus, and evaluation uses the standard test set.

Models and Retrieval: We use Qwen2.5-instruct-

7b (Yang et al., 2024b) as our base model, and use the BM25 (Robertson and Zaragoza, 2009) as reproduce RAG methods retriever.

Metrics: We use Avg@5, Cons@5, and Pass@5, representing average accuracy over 5 generations, majority-vote accuracy, and the rate of including at least one correct answer.

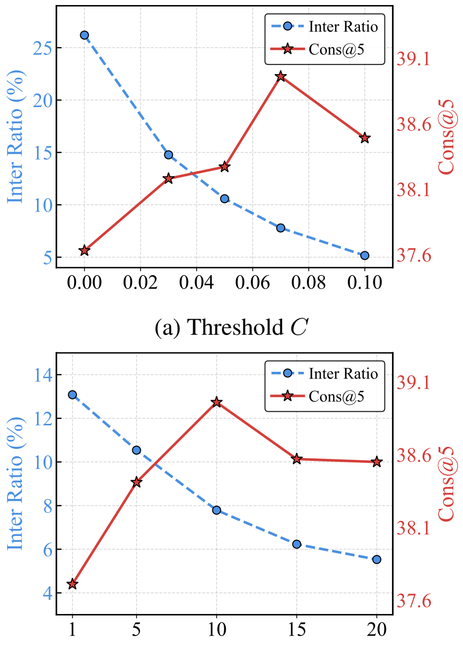
Baselines: We compare S2K with representative methods across four categories: prompting, RAG, post-training, and domain-specific pretraining. Prompting includes 0/1/3/5-shot settings. RAG baselines cover standard RAG, Self-RAG (Asai et al., 2023), and Self-Ask (Press et al., 2023). Post-training includes SFT, PPO, and GRPO under consistent conditions. We also compare with domain-specific pretrained models, including BioMistral (Labrak et al., 2024), MMed-Llama-3-8B (Qiu et al., 2024), and OpenBioLLM-8B (Ankit Pal, 2024) for medicine; Saul-7B (Colombo et al., 2024), LawChat (Cheng et al., 2024b), and Lawyer-LLaMA-13B (Huang et al., 2023) for law; and finance-Llama3-8B (Cheng et al., 2024a), xunayuan-6B-chat (Zhang and Yang, 2023), and CFGPT (Li et al., 2024) for finance.

More implementation details, including hyperparameters and baselines, are provided in the Appendix A.

4.2 Main Result

We evaluate S2K from two perspectives. At the algorithm level, we reproduce and compare representative QA enhancement methods, including prompting strategies, training techniques, and retrieval-augmented generation, under identical settings for fair comparison. At the model level, we directly compare with open-source domain-specific pretrained models to demonstrate the effectiveness of our approach in realistic deployment scenarios.

S2K proves to be the most effective method for enhancing DSQA. As shown in Table 1, it consistently delivers significant performance gains across all three domains compared to the raw LLM, demonstrating strong generalization capabilities. Moreover, it outperforms all other QA enhancement strategies we benchmarked. Notably, methods that inject domain knowledge into the model's context (e.g., Few-Shot and RAG) generally underperform, suggesting that in knowledge-intensive tasks, especially those requiring complex reasoning, embedding knowledge directly into model parameters is a more promising approach.



(a) Threshold C

(b) Window W

Figure 4: Effect of Threshold C and Window W in Knowledge Self-Selection.

S2K achieves competitive performance with domain-pretrained models at a significantly lower training cost. As shown in Figure 3, while domain-specific pretraining typically requires hundreds of billions of tokens, S2K uses two to three orders of magnitude less data (e.g., only 0.04B tokens for the medical domain), yet still matches or even surpasses their performance across all three domains. This highlights the effectiveness of fusing internal parametric knowledge with external domain knowledge, which will become increasingly valuable as LLMs continue to improve in their internal knowledge in the future.

4.3 Analysis Experiments

4.3.1 Threshold C in Knowledge Selection

As shown in Equation 3, we introduce threshold C in internal-external knowledge fusion to encourage more cautious selection of internal knowledge. As illustrated in Figure 4a, we analyze the effect of C on both the proportion of internal knowledge in the fused data and the model's performance, increasing C from 0 to 0.1 reduces the proportion of selected internal tokens from 26.20% to 5.16%, aligning with the self-selection mechanism defined in Equation 3. Interestingly, model performance first improves and then declines as C increases, peaking at $C = 0.07$. This suggests that an overly high proportion of internal knowledge may lead

Sampling	Avg@5	Cons@5	Pass@5
Random	32.6	35.0	44.6
Relevance-based	38.6	43.4	77.1

Table 2: Effect of sampling strategies on reasoning data generation.

Acc	Fmt		Metrics		
Correct	*Correct*	*EA*	Avg@5	Cons@5	Pass@5
1	-	-	34.9	35.7	61.3
1	1	-0.5	35.6	36.7	54.9
5	1	-0.5	**38.6**	**43.4**	**77.1**

Table 3: Comparison of reward schemes. While Acc means Accuracy reward, Fmt means format reward and *EA* means extra-answer penalty.

to overconfidence. Conversely, when the internal knowledge proportion is too low, the fusion reduces to relying solely on external knowledge, thereby neglecting the utility of useful internal knowledge.

4.3.2 Window Width W in Knowledge Fusion

To mitigate greedy selection behavior when fusing knowledge, we introduce a window size parameter W in Equation 3. The model selects internal knowledge based on the average log-probability over a window of W tokens, instead of a single token level. As shown in Figure 4b, W increases from 1 to 20, the proportion of selected internal tokens steadily decreases. This indicates that the window mechanism effectively alleviates greedy selection. Correspondingly, model performance first improves and then degrades, peaking at $W = 10$, suggests that a larger window smooths locally confident but potentially incorrect predictions, encouraging the model to be more cautious in selecting internal knowledge, but an excessively large window may overly suppress internal knowledge, causing the model to rely entirely on external knowledge.

4.3.3 Relevance-based sampling of Reasoning Data Generation

As mentioned in Section 3.1.3, we hypothesize that complex reasoning tasks require the integration of multiple relevant knowledge points. To better simulate realistic reasoning scenarios, we introduce a relevance-based sampling strategy during the generation of reasoning data. In this section, we quantitatively compare the effects of random and relevance-based sampling on model performance. The results in Table 2 show that relevance-based sampling significantly improves model performance, supporting the validity of our hypothesis.

Setting	Avg@5	Cons@5	Pass@5
Raw LLM	33.5	38.3	67.6
Std. SFT & Ext. Data	33.5	36.8	68.7
Sel. SFT & Ext. Data	34.2	37.9	73.1
Sel. SFT & Fus. Data	35.1	39.6	75.9
Only GRPO	36.1	36.4	61.4
S2K	38.6	43.4	77.1

Table 4: Ablation study. 🟥 Internal–External Fusion fine-grained ablation; 🟦 End to End ablation; (Abbreviations: Std. SFT=Standard SFT; Sel. SFT=Selective SFT; Ext. Data=External Training Data; Fus. DT=Fusion Training Data)

4.3.4 Reward Function Analysis

We use GRPO with accuracy and format rewards to boost QA performance in real-world, domain-specific settings. We compare three reward schemes: *(1) Answer Only:* binary reward for answer correctness; *(2) Answer + Format:* combined reward for correctness and formatting; and *(3) Enhanced Answer + Format:* combined reward with stronger Answer incentives.

As shown in Table 3, the answer only reward can lead to formatting issues that degrade overall performance. Adding a formatting reward significantly improves structural consistency, although it lags behind in terms of correctness. By contrast, increasing the answer reward while still incorporating the formatting reward achieves the best results. Therefore, we ultimately select the third reward scheme as the reward during the Reasoning-Enhanced Training.

4.4 Ablation Study

To further validate the contribution of each component in S2K, we conduct a detailed ablation study covering internal-external knowledge fusion, reinforcement learning, and end-to-end training. As shown in Table 4, during the first-stage training, our proposed Selective SFT (Section 3.2) outperforms standard SFT, and the fusion of internal and external knowledge (Section 3.1.2) leads to better performance than using external knowledge alone, demonstrating the effectiveness of both components. Furthermore, compared to the final model trained with the full two-stage pipeline (Avg@5: 38.7), models trained with only Selective SFT (35.2) or only GRPO (36.5) exhibit inferior performance, highlighting the importance and effectiveness of our overall training strategy.

Type	Content
Question	What are the key functional differences between M1 and M2 macrophages in their metabolism of arginine during the immune response to helminths?
Document	...A major difference between M1 and M2 macrophages is...Whereas **M1 macrophages express iNOS**, which produces the potent intracellular microbicide nitric oxide (NO), **M2 macrophages express arginase-1**, which produces ornithine and proline from arginine...
External Answer	...**M1 macrophages express iNOS**, which produces nitric oxide (NO)...**M2 macrophages express arginase-1**. Arginase-1 breaks down arginine into ornithine and proline... M1 is usually associated with Th2 cells ✗ and promotes tissue repair and anti-inflammatory responses. M2 is linked to Th1 cells ✗ and promotes defense...
Fusion Answer	...**M1 macrophages express iNOS**, which produces nitric oxide (NO)...M1 macrophages are typically associated with the Th1 response ✓... **M2 macrophages express arginase-1.** Arginase-1 breaks down arginine into ornithine and proline... M2 macrophages are linked to the Th2 response ✓...

Table 5: Knowledge comparison between different answer sources and the fusion result. The original document accurately distinguishes the metabolic roles of M1 and M2 macrophages. External data reiterates some facts but introduces significant errors, such as wrongly linking M1 macrophages to Th2 responses. Our fusion method effectively corrects these inaccuracies while preserving useful complementary details from the external source.

4.5 Case Study

In this section, we present a real case in Table 5, to demonstrate how our fusion mechanism works. The original document describes the functional differences between M1 and M2 macrophages in arginine metabolism, while the external, though containing some relevant facts, introduces notable knowledge errors. Our fusion answer successfully identifies and corrects these errors while retaining complementary details from the external source, resulting in a more complete and accurate answer.

5 Related Work

Domain-Specific Question Answering: Domain-Specific QA (Zhang et al., 2024b; Wang et al., 2024; Siriwardhana et al., 2023) involves leveraging LLMs to accurately understand and respond to user queries in specialized fields such as medicine, law, and finance. Despite recent advancements, LLMs still exhibit noticeable performance gaps in DSQA tasks (Yang et al., 2023; Mao et al., 2024; Sharma et al., 2024; Yue, 2025). This shortfall is primarily due to two key challenges. First, general-purpose LLMs often lack sufficient domain-specific knowledge (Mao et al., 2024; Bhushan et al., 2025). Second, hallucinations (Ji et al., 2023; Sultania et al., 2024; Bhushan et al., 2025) remain a major concern, while LLMs can generate fluent and coherent responses, but may be factually incorrect or misaligned with the original sources.

Retrieval-Augmented Generation: RAG (Guu et al., 2020; Lewis et al., 2020; Izacard et al., 2022; Nakano et al., 2021; Asai et al., 2023; Ma et al., 2023; Yu et al., 2024; Shi et al., 2024) enhances LLMs by incorporating external domain-specific knowledge, to mitigate hallucinations and improve performance in DSQA tasks (e.g., Self-RAG (Asai et al., 2023) is capable of dynamically determining whether domain-specific knowledge needs to be retrieved based on the query context, while Self-Ask (Press et al., 2023) uses search engines for sub-questions). However, it suffers from conflicting internal and external domain knowledge (Xu et al., 2024; Zhang et al., 2024a; Xie et al., 2024).

Continued Training Domain Adaptation: Continued training (Labrak et al., 2024; Qiu et al., 2024; Zhang et al., 2025; Mecklenburg et al., 2024) aims to inject domain-specific knowledge into LLMs to compensate for their lack of specialized expertise. This strategy can be broadly divided into two main approaches: pre-training (Qiu et al., 2024; Shu et al., 2024; Li et al., 2024; Chen et al., 2023) adaptation, which fine-tunes LLMs on domain-specific corpora to help them internalize expert knowledge (e.g., BioMistral (Labrak et al., 2024)); and post-training (Zhang et al., 2025; Mecklenburg et al., 2024), which involves fine-tuning LLMs using QA pairs derived from domain knowledge. However, continued training often encounters hurdles in effectively enabling LLMs to extract the acquired knowledge during the inference phase (Zhang et al., 2025; Ibrahim et al.; Ovadia et al., 2024).

6 Conclusion

To address challenges in DSQA, we propose S2K, an efficient framework designed to enhance the performance of LLMs in long-tail domains. In vertical domains where no readily available QA datasets exist, S2K enables effective transfer and generalization of QA capabilities using only raw corpora. Experiments across multiple representative vertical domain results demonstrate its effectiveness.

7 Limitation

Although S2K demonstrates strong performance across various domain-specific scenarios, there remains room for further improvement. At present, the method primarily focuses on modeling static domain knowledge and has not been specifically optimized for rapidly evolving or real-time information. In the future, we plan to integrate RAG techniques to enhance the system's adaptability to dynamic knowledge while maintaining broad coverage.

References

Malaikannan Sankarasubbu Ankit Pal. 2024. Openbiollms: Advancing open-source large language models for healthcare and life sciences. https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Kushagra Bhushan, Yatin Nandwani, Dinesh Khandelwal, Sonam Gupta, Gaurav Pandey, Dinesh Raghu, and Sachindra Joshi. 2025. Systematic knowledge injection into large language models via diverse augmentation for domain-specific RAG. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5922–5943, Albuquerque, New Mexico. Association for Computational Linguistics.

Steven Bird. 2006. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 interactive presentation sessions*, pages 69–72.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Wei Chen, Qiushi Wang, Zefei Long, Xianyin Zhang, Zhongtian Lu, Bingxuan Li, Siyuan Wang, Jiarong Xu, Xiang Bai, Xuanjing Huang, et al. 2023. Discfinllm: A chinese financial large language model based on multiple experts fine-tuning. *CoRR*.

Daixuan Cheng, Yuxian Gu, Shaohan Huang, Junyu Bi, Minlie Huang, and Furu Wei. 2024a. Instruction pre-training: Language models are supervised multitask learners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2529–2550.

Daixuan Cheng, Shaohan Huang, and Furu Wei. 2024b. Adapting large language models via reading comprehension. In *The Twelfth International Conference on Learning Representations*.

Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, Andre FT Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia Morgado, et al. 2024. Saullm-7b: A pioneering large language model for law. *arXiv preprint arXiv:2403.03883*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

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.

Bolei He, Nuo Chen, Xinran He, Lingyong Yan, Zhenkai Wei, Jinchang Luo, and Zhen-Hua Ling. 2024. Retrieving, rethinking and revising: The chain-of-verification can improve retrieval augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10371–10393, Miami, Florida, USA. Association for Computational Linguistics.

Quzhe Huang, Mingxu Tao, Chen Zhang, Zhenwei An, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. 2023. Lawyer llama technical report. *arXiv preprint arXiv:2305.15062*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Adam Ibrahim, Benjamin Thérien, Kshitij Gupta, Mats Leon Richter, Quentin Gregory Anthony, Eugene Belilovsky, Timothée Lesort, and Irina Rish. Simple and scalable strategies to continually pre-train large language models. *Transactions on Machine Learning Research*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

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).

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. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5848–5864, Bangkok, Thailand. Association for Computational Linguistics.

Patrick 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*, volume 33, pages 9459–9474. Curran Associates, Inc.

Jiangtong Li, Yang Lei, Yuxuan Bian, Dawei Cheng, Zhijun Ding, and Changjun Jiang. 2024. Ra-cfgpt: Chinese financial assistant with retrieval-augmented large language model. *Frontiers of Computer Science*, 18(5):185350.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting in retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5303–5315, Singapore. Association for Computational Linguistics.

Kelong Mao, Zheng Liu, Hongjin Qian, Fengran Mo, Chenlong Deng, and Zhicheng Dou. 2024. Ragstudio: Towards in-domain adaptation of retrieval augmented generation through self-alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 725–735.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 1906. Association for Computational Linguistics.

Nick Mecklenburg, Yiyou Lin, Xiaoxiao Li, Daniel Holstein, Leonardo Nunes, Sara Malvar, Bruno Silva, Ranveer Chandra, Vijay Aski, Pavan Kumar Reddy Yannam, Tolga Aktas, and Todd Hendry. 2024. Injecting new knowledge into large language models via supervised fine-tuning. *Preprint*, arXiv:2404.00213.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. 2024. Fine-tuning or retrieval? comparing knowledge injection in llms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 237–250.

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

Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Towards building multilingual language model for medicine. *Nature Communications*, 15(1):8384.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*.

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

Pasi Shailendra, Rudra Chandra Ghosh, Rajdeep Kumar, and Nitin Sharma. 2024. Survey of large language models for answering questions across various fields. In *2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS)*, volume 1, pages 520–527.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Sanat Sharma, David Seunghyun Yoon, Franck Dernoncourt, Dewang Sultania, Karishma Bagga, Mengjiao Zhang, Trung Bui, and Varun Kotte. 2024. Retrieval augmented generation for domain-specific question answering. *CoRR*.

Zhengliang Shi, Shuo Zhang, Weiwei Sun, Shen Gao, Pengjie Ren, Zhumin Chen, and Zhaochun Ren. 2024. Generate-then-ground in retrieval-augmented generation for multi-hop question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7339–7353.

Dong Shu, Haoran Zhao, Xukun Liu, David Demeter, Mengnan Du, and Yongfeng Zhang. 2024. Lawllm: Law large language model for the us legal system. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 4882–4889. ACM.

Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana, and Suranga Nanayakkara. 2023. Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering. *Transactions of the Association for Computational Linguistics*, 11:1–17.

Dewang Sultania, Zhaoyu Lu, Twisha Naik, Franck Dernoncourt, David Seunghyun Yoon, Sanat Sharma, Trung Bui, Ashok Gupta, Tushar Vatsa, Suhas Suresha, Ishita Verma, Vibha Belavadi, Cheng Chen, and Michael Friedrich. 2024. Domain-specific question answering with hybrid search. *Preprint*, arXiv:2412.03736.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Shuting Wang, Jiongnan Liu, Shiren Song, Jiehan Cheng, Yuqi Fu, Peidong Guo, Kun Fang, Yutao Zhu, and Zhicheng Dou. 2024. Domainrag: A chinese benchmark for evaluating domain-specific retrieval-augmented generation. *CoRR*.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, Miami, Florida, USA. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024b. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Fangkai Yang, Pu Zhao, Zezhong Wang, Lu Wang, Bo Qiao, Jue Zhang, Mohit Garg, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. 2023. Empower large language model to perform better on industrial domain-specific question answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 294–312.

Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. 2023. Cognitive mirage: A review of hallucinations in large language models. *arXiv preprint arXiv:2309.06794*.

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 *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Murong Yue. 2025. A survey of large language model agents for question answering. *arXiv preprint arXiv:2503.19213*.

Hao Zhang, Yuyang Zhang, Xiaoguang Li, Wenxuan Shi, Haonan Xu, Huanshuo Liu, Yasheng Wang, Lifeng Shang, Qun Liu, Yong Liu, et al. 2024a. Evaluating the external and parametric knowledge fusion of large language models. *CoRR*.

Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Yipeng Zhang, Haitao Mi, and Helen Meng. 2025. Self-tuning: Instructing llms to effectively acquire new knowledge through self-teaching. *Preprint*, arXiv:2406.06326.

Xuanyu Zhang and Qing Yang. 2023. Xuanyuan 2.0: A large chinese financial chat model with hundreds of billions parameters. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, pages 4435–4439.

Yichi Zhang, Zhuo Chen, Yin Fang, Yanxi Lu, Li Fangming, Wen Zhang, and Huajun Chen. 2024b. Knowledgeable preference alignment for LLMs in domain-specific question answering. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 891–904, Bangkok, Thailand. Association for Computational Linguistics.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Jec-qa: A legal-domain question answering dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9701–9708.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. Agieval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314.

A Appendix

A.1 Visualization of token-level prediction probabilities

Figure 1 illustrates the importance of internal parametric knowledge using a schematic example, while Figure 5 presents a real-world case. We randomly sample document chunks from a medical document and feed them into the LLM. Based on the model's output logits, we compute token probabilities and visualize the top 32 tokens with the highest confidence. The results show that even when provided with external domain documents, the model correctly predicts a substantial portion of tokens with high confidence. This indicates that the LLM has already acquired part of this domain knowledge during pretraining.

A.2 Implementation Details

This section provides a detailed overview of the experimental details, including data scales for training and evaluation, hyperparameter configurations, and analysis experiments, to ensure the reproducibility and rigor of our results.

A.2.1 Datasets

We first extract meta knowledge from raw domain-specific corpora. For each meta knowledge instance, we generate internal-external fused data. Additionally, we sample multiple meta knowledge entries to construct complex reasoning examples. Experiments are conducted in three domains: medicine, law, and finance. The number of samples for each data type in each domain is summarized in Table 6.

Domain	$\mathcal{D}_{meta}$	$\mathcal{D}_{fusion}$	$\mathcal{D}_{reason}$	$\mathcal{D}_{eval}$
Medicine	41760	41760	3492	1273
Law	15332	15332	4297	1000
Finance	29789	29789	1505	7123

Table 6: Number of samples datasets: where $\mathcal{D}_{meta}$ means Meta Knowledge, $\mathcal{D}_{fusion}$ means Fusion Knowledge number, $\mathcal{D}_{reason}$ means Reasoning Knowledge, $\mathcal{D}_{eval}$ means evaluate samples numbers.

A.2.2 Hyperparameter

As described in Section 3.2, our proposed Selective SFT introduces a weighting factor to the standard SFT loss, with weights ranging from 0 to 1. As a result, the overall loss in Selective SFT is smaller than that of standard SFT. To compensate and enhance training effectiveness, we increase the learning rate accordingly. Table 7 presents the detailed hyperparameter settings for Selective SFT.

Hyperparameter	Value
Finetuning Type	lora
Lora Rank	8
Batch Size	32
Learning Rate	1e-3
Number of Epochs	1.0
LR Scheduler	cosine
Warm-up Ratio	0.1

Table 7: Hyperparamters of Selective SFT.

In addition, Table 8 provides the detailed hyperparameter settings used in the GRPO stage. For fair comparison, the reinforcement learning baselines are configured with the same hyperparameters.

Hyperparameter	Value
Number of Epochs	2
Learning Rate	5e-6
Sequence Length	4096
Warm-up Ratio	0.1
Global Batch Size	1
Rollout Batch Size	8
Max Prompt Length	512
Max Response Length	2048
KL Coefficient	0.04
Checkpoint Strategy	step
Random Seed	42
Temperature	0.9
Top-p	1.0
Max grad norm	0.1

Table 8: Hyperparameters of Reinforce Learning.

A.2.3 Metric

We evaluate model performance using three metrics computed over $k = 5$ generated answers per question: Avg@5, Cons@5, and Pass@5. Given a set of N questions, for each question i we denote the set of generated answers as $a_{i1}, a_{i2}, \ldots, a_{i5}$ and their correctness as binary indicators $y_{i1}, y_{i2}, \ldots, y_{i5}$ where $y_{ij} = 1$ if a_{ij} is correct, otherwise 0.

Avg@5 measures the average accuracy across all 5 generations:

$$\text{Avg@5} = \frac{1}{5N} \sum_{i=1}^{N} \sum_{j=1}^{5} y_{ij} \qquad (7)$$

Cons@5 evaluates the correctness of the majority-voted answer among the 5 generations:

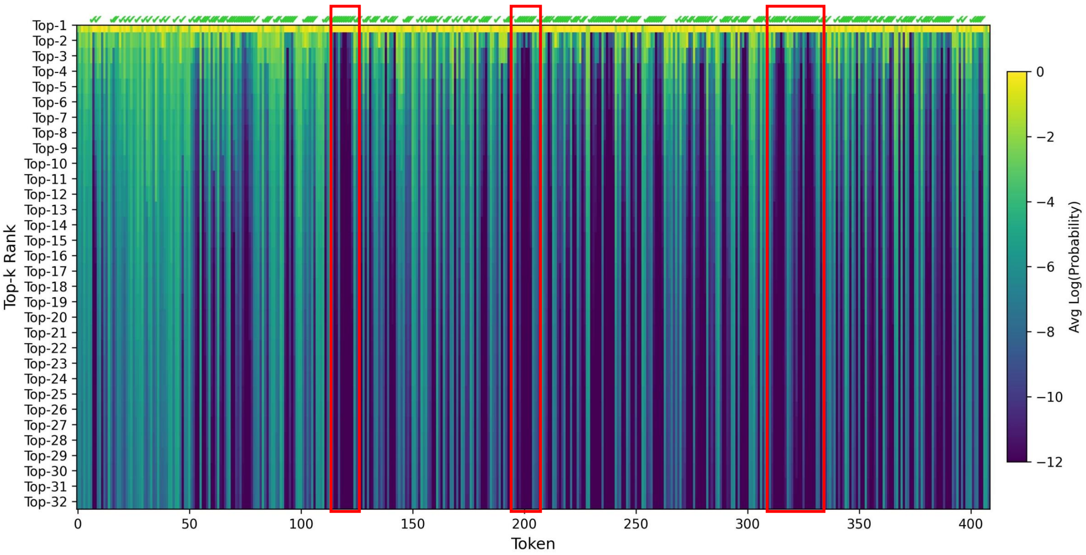


Figure 5: A real example of token-level prediction probabilities. The horizontal axis represents the token positions in a domain-specific document, and the vertical axis shows the top-32 tokens ranked by predicted probability. Green check marks at the top indicate tokens correctly predicted by the model. A greater vertical spread of green marks suggests more dispersed probabilities and lower model confidence. In contrast, concentrated predictions with high-ranked correct tokens indicate strong confidence, implying that the model has already internalized the corresponding domain knowledge.

$$\text{Cons@5} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left(\text{major}(a_{i1},\ldots,a_{i5}) = a_i^{\text{gold}}\right) \quad (8)$$

where major$(\cdot)$ returns the most frequent answer among the 5 generations, and a_i^{gold} is the correct answer for question i. $\mathbb{I}(\cdot)$ is the indicator function, which returns 1 if the condition is true and 0 otherwise.

Pass@5 measures whether at least one of the 5 generations is correct:

$$\text{Pass@5} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left(\sum_{j=1}^{5} y_{ij} \geq 1\right) \quad (9)$$

A.3 Baseline Reproduction Details

In this section, we provide a detailed description of the reproduction process for other methods to demonstrate the reproducibility and fairness of the experimental comparisons.

A.3.1 Few-Shot

In Table 1, we include 0/1/3/5-shot prompting as baselines. The zero-shot setting corresponds to the raw LLM, while the 1/3/5-shot prompts are randomly sampled from each dataset's official training set. For each test sample, the prompts are independently sampled, with a fixed random seed to ensure reproducibility.

A.3.2 Hyperparameter Settings for Reinforcement Learning Methods

To ensure reproducibility and fair comparison, we closely followed standard implementations and platform-recommended values when reproducing baseline reinforcement learning methods. Table 8 summarizes the key hyperparameters. The configuration was applied consistently across all PPO and GRPO training runs. All experiments were conducted under the same hardware environment and data preprocessing pipeline.

A.3.3 Hyperparameter Settings for RAG with BM25 Retrieval

For experiments involving RAG, we adopt a traditional BM25-based retriever to collect candidate documents, followed by a reranking stage to refine the top selections. The key parameters used in both retrieval and reranking stages are summarized below. Retrieval is performed using a batch-based setup with FP16 precision enabled for improved efficiency. Reranking similarly operates in batches, with truncated input lengths to balance context and computational cost.Table 9 summarizes the key hyperparameters used for RAG.

Hyperparameter	**Value**	**Description**
Random Seed	2024	Seed for reproducibility in retrieval and reranking.
Retrieval Top-k	5	Number of top documents retrieved per query.
Retrieval Batch Size	256	Number of queries processed in parallel during retrieval.
Retrieval FP16	True	Use half-precision (FP16) for retrieval computations.
Retrieval Max Query Length	128	Max token length for each query input.
Rerank Top-k	5	Number of documents reranked per query after initial retrieval.
Rerank Max Length	512	Max token length for concatenated query-document input to reranker.
Rerank Batch Size	256	Number of samples reranked in parallel.
Rerank FP16	True	Use FP16 precision for reranking to reduce memory usage.

Table 9: Hyperparameter settings for RAG pipeline with BM25-based retrieval and reranking.

A.4 Structured Reasoning Examples in QA Generation

Table 10 presents representative examples of the three structured reasoning types—**deductive**, **inductive**, and **case-based**—used in our QA pair generation framework. These examples were constructed to reflect clinically relevant diagnostic and management scenarios, enabling the large language model (LLM) to generate complex question–answer pairs guided by distinct logical paradigms.

- The **deductive** example demonstrates reasoning from a general diagnostic framework (DSM coding rules) to a specific clinical case involving substance-induced depressive disorder.
- The **inductive** example illustrates how generalizable conclusions can be drawn from specific patient findings that align with the Brighton diagnostic criteria for Guillain-Barré syndrome.
- The **case-based** example applies analogical reasoning to a surgical decision-making scenario, where guideline-informed management is inferred based on patient characteristics and clinical history.

These examples serve to clarify how our structured prompting strategy supports logical diversity and clinical fidelity in synthetic QA data creation, and they provide concrete evidence of how different reasoning pathways are operationalized in practice.

A.5 Prompts

We generate meta knowledge and reasoning data by prompting DeepSeek-v3 and GPT-4o. This section provides detailed prompt templates. Table 11 presents the prompt used for meta knowledge generation, while Tables 12 to 14 show the prompts used for generating inductive, deductive, and case-based reasoning data, respectively.

Category	Content
Input	1. What are the primary functions of visceral sensory fibers associated with parasympathetic fibers in the nervous system? (Document:For example, visceral sensory fibers from the heart may enter at...) 2. What are the roles of the parasympathetic and sympathetic nerve supplies in modulating motor functions of the small intestine and colon? (Document:The myenteric plexus regulates smooth-muscle function through...) ... 9. What are the functional roles of the enteric system in the gastrointestinal tract, and how does it interact with the sympathetic and parasympathetic nervous systems? (Document:These branches contain preganglionic parasympathetic fibers...) 10. What are the anatomical structures and spinal segments involved in the sympathetic, parasympathetic, and sensory innervation of the uterovaginal plexus? (Document:Uterovaginal plexus Frankenhäuser...)
Deductive	A patient presents with chronic abdominal pain and irregular bowel movements. Upon further investigation, it is suspected that the enteric nervous system is not properly regulating gastrointestinal activities. The physician hypothesizes that the issue may involve improper modulation by extrinsic parasympathetic fibers. Based on the anatomical and functional organization of the parasympathetic nervous system, which of the following spinal cord levels would most likely contain the preganglionic parasympathetic neurons involved in regulating the distal colon? A. T1–T12 B. L1–L2 C. S2–S4 D. C1–C7 Correct Answer: C
Inductive	A 45-year-old woman presents with chronic pelvic pain and dysmenorrhea due to endometriosis. Her physician considers performing a presacral neurectomy to interrupt the sensory fibers from the uterus and cervix. Based on the anatomical and functional innervation of the uterovaginal plexus, which of the following statements is correct regarding this procedure? A. The procedure will effectively relieve pain originating from the ovary because the ovarian plexus is part of the superior hypogastric plexus. B. The procedure carries a risk of bladder and bowel dysfunction because the superior hypogastric plexus contains efferent sympathetic and parasympathetic fibers. C. The procedure targets sensory fibers entering the spinal cord at levels S2–S4, which are responsible for uterine pain. D. The procedure is unlikely to affect peristalsis in the distal colon because the enteric system operates independently of extrinsic innervation. Correct Answer: B
Case-based	A 45-year-old woman presents to her physician with chronic pelvic pain and dysmenorrhea. Upon further evaluation, she is diagnosed with endometriosis. The physician considers surgical intervention to alleviate her symptoms by targeting the sensory innervation of the uterus and cervix. Based on the anatomical and functional information provided, which of the following surgical approaches would most likely interrupt the sensory fibers responsible for her pain while minimizing the risk of bladder or bowel dysfunction? A. Resection of the ovarian plexus B. Presacral neurectomy (resection of the superior hypogastric plexus) C. Blockade of the pudendal nerve D. Resection of a portion of the uterosacral ligaments Correct Answer: D

Table 10: Representative Examples of Structured Reasoning Types Used in QA Pair Generation. Each case illustrates a distinct reasoning paradigm—deductive, inductive, or case-based.

Prompt #1: Meta Knowledge Generation

~~~
You are a professional question-generation expert with a focus on academic and technical texts.

## Task:
Carefully read the provided document chunk and generate **exactly one knowledge-based, specific, and self-contained question**. The question must:
1. Be directly answerable using only the content from the chunk.
2. Reflect representative or meaningful knowledge contained in the chunk — not superficial, vague, or structural elements.
3. Be expressed in formal, academic language, precise and clear.

## Rules:
1. The question must be fully self-contained and understandable without access to the original chunk.
2. Do **NOT** use context-dependent phrases like: ”as described in the text”, ”according to the passage”, ”in the document”, ”from the chunk”
3. Add necessary information to the question to ensure that it can be independently understood. (Bad Case: What are the symptoms described in the text? Good Case: What are the typical symptoms of generalized anxiety disorder?)
4. If the chunk lacks sufficient knowledge content or contains only general statements, structural formatting, or introductory language, return the JSON format with an **empty question string**.
5. Avoid vague or incomplete questions like ”What does X refer to?”
6. If necessary, add contextual qualifiers (e.g., domain, subject, scope) to the question to ensure it is fully understandable without seeing the original chunk.
7. Favor questions that involve comparisons, causes, functions, conditions, or processes over basic definitional questions.
8. If possible, vary the question style (e.g., what, why, how), but keep it answerable solely from the chunk.

## Output Format:
Only respond in this strict JSON format, without any extra text, markdown, or commentary:

``` json
{{
”question”: ”Your single knowledge-based question here — or an empty string if no meaningful question can be asked.”
}}
```

## Document:
{article_text}
~~~

Table 11: Prompt Design for Meta Knowledge Generation

Prompt #2: Inductive Reasoning Data Generation

```
### General Instruction
You are an advanced question generation model that aims to generate case questions that require
inductive reasoning based on multiple instances or observations in the text. Your task is to
generate a question that requires synthesizing information from the provided factual questions
and their corresponding texts. The question must be complete and understandable without requiring
external information.

### Reasoning Type Requirement: Inductive reasoning
A "Instruction" questions involve presenting a realistic scenario where the rules must come
directly from the text (e.g., definitions, theorems, taxonomies). The scenario should be coherent
and plausible in the context of the given information. Not all input information needs to be used;
select the most relevant parts to construct a meaningful question.

### Question Type Requirement: Multiple choice
The generated questions should be presented in the form of multiple-choice questions with **four
options (A, B, C, D)**, only one of which is correct. The correct answer can directly match the
inevitable conclusion in the rule, and the statement should be clear and avoid vague words. Make
sure the distractors seem reasonable, but are obviously different from the correct answer.

### Additional Notes
1. Use clear and concise language to present the scenario.
2. Avoid unnecessary complexity, but ensure the question requires reasoning beyond direct
retrieval.
3. Make sure the question is self-contained and understandable without additional context, that
is, you can understand without using the content in the text.

### Example Format:
Input:
1. What role do natural killer (NK) cells play in immunosurveillance? (Text: Natural killer
(NK) cells play a critical role in the immune response against tumors by killing cancer cells
through perforin-mediated cytotoxicity, which is essential for immunosurveillance in the body.
This process helps to limit tumor progression, making NK cells important in the study of cancer
prognosis.)
2. Which cluster of differentiation marker is used to identify natural killer (NK) cells in
tissue samples? (Text: CD56 is a surface marker specific to natural killer (NK) cells and is used
to identify and isolate these cells in tissue samples, such as those from resected lung cancer
specimens. Therefore, CD56 is the appropriate cluster of diff rentiation marker to study tumor
infiltration by NK cells in cancer research.)

Output:
In a suburban town in Virginia, epidemiologists are alarmed by the increasing number of squamous
cell lung cancer cases. Further investigation reveals that most people in the area work in a
glass factory, the regions main source of employment. A researcher is interested in studying the
role of immunosurveillance in the pathogenesis of this lung cancer. He postulates that tumor
infiltration by natural killer (NK) cells has a better prognosis since they play a major role
in immunosurveillance. NK cells also kill tumor cells by the perforin-mediated destruction of
cancerous cells. The researcher is interested in studying tumor infiltration by NK cells in the
resected specimen from patients within the cohort who have been diagnosed with stage 1 lung cancer.
Which of the following cluster of differentiation markers will he need to use to identify these
cells in the resected specimens?
A. CD20
B. CD3
C. CD34
D. CD56
Correct Answer: D

### Input:
{meta_knowledge_from_sampling}

Now start generating one question based on the given input.
```

Table 12: Prompt Design for Inductive Reasoning Data Generation

Prompt #3: Deductive Reasoning Data Generation

```
### General Instruction
You are an advanced question generation model that aims to generate case questions that require
deductive reasoning based on the knowledge points in the question and the general rules or
definitions in the text. You need to extract clear rules from the text and design a realistic
scenario that requires users to solve the problem through logical deduction from general to specific.

### Reasoning Type Requirement: Deductive reasoning
A ”deductive” question involves presenting a realistic scenario where information from the provided
texts must be applied to diagnose, explain, or solve a specific problem. The scenario should be
coherent and plausible within the context of the given information. Not all input information
needs to be used; select the most relevant parts to construct a meaningful question.

### Question Type Requirement: Multiple Choice The generated question should be presented
as a multiple-choice question with **four options (A, B, C, D)**, where only one option is correct.
Ensure the distractors are plausible but clearly distinguishable from the correct answer. The user
should be able to choose the correct answer by synthesizing information from the provided factual
questions and texts.

### Additional Notes
1. Use clear and concise language to present the scenario.
2. Avoid unnecessary complexity, but ensure the question requires reasoning beyond direct
retrieval.
3. Make sure the question is self-contained and understandable without additional context.

### Example Format:

Input:
1. What are the four primary features of tetralogy of Fallot? (Text: Tetralogy of Fallot is a
congenital heart defect characterized by four primary features: ventricular septal defect (VSD),
pulmonary stenosis, right ventricular hypertrophy (RVH), and overriding aorta. These abnormalities
can lead to cyanosis, particularly during episodes of increased oxygen demand, such as feeding or
crying.)
2. Why is right axis deviation a common finding on the electrocardiogram (ECG) of patients
with tetralogy of Fallot? (Text: In patients with tetralogy of Fallot, the electrocardiogram
(ECG) commonly shows right axis deviation due to the right ventricular hypertrophy (RVH) that
develops as a result of the obstruction to blood flow through the pulmonary valve. This feature
is characteristic of the condition and helps to differentiate it from other congenital heart
defects.)

Output:
A 6-month-old girl presents with cyanosis of the lips during feeding. The father reports that the
child has similar brief episodes during activity. Physical examination reveals that the child’s
lips and fingers have cyanosis induced by crying during ear examination. Based on the diagnostic
criteria for tetralogy of Fallot, which of the following features is most likely to be shown on
the child’s electrocardiogram?
A. Left ventricular hypertrophy
B. ST segment depression
C. Widened QRS complex
D. Right axis deviation
Correct Answer: D

### Input: {meta_knowledge_from_sampling}

Now start generating one question based on the given input.
```

Table 13: Prompt Design for Deductive Reasoning Data Generation

Prompt #4: Case Reasoning Data Generation

```
### General Instruction
You are an advanced question generation model designed to create comprehensive reasoning questions based on factual questions and their corresponding text passages. Your task is to generate a question that requires synthesizing information from the provided factual questions and their corresponding texts. The question must be complete and understandable without requiring external information.

### Reasoning Type Requirement: Case
A "Case" question involves presenting a realistic scenario where information from the provided texts must be applied to diagnose, explain, or solve a specific problem. The scenario should be coherent and plausible within the context of the given information. Not all input information needs to be used; select the most relevant parts to construct a meaningful question.

### Question Type Requirement: Long form
The generated question should be presented as a long form. The user should be able to answer by synthesizing information from the provided factual questions and texts.

### Additional Notes
1. Use clear and concise language to present the scenario.
2. Avoid unnecessary complexity, but ensure the question requires reasoning beyond direct retrieval.
3. Make sure the question is self-contained and understandable without additional context.

### Example Format:

Input:
1. What is the infectious form of the prion protein associated with scrapie called? (Text: The infectious form of the prion protein associated with scrapie is PrPSc, which is misfolded and can induce other proteins to misfold as well.)
2. What is the role of myoglobin in muscle cells concerning oxygen management? (Text: Myoglobin serves as an oxygen storage molecule in muscle cells, allowing oxygen to be available during periods of intense activity.)

Output:
A 55-year-old sheep farmer reports that several of his sheep are exhibiting unusual symptoms such as tremors, lack of coordination, and intense itching that leads to wool loss. Additionally, he mentions feeling tired quickly during routine tasks such as herding the sheep. The farmer is concerned that the symptoms may be related to some infectious agent present on the farm. Based on the symptoms described and the information provided, what could be the cause of the sheep's condition?
Correct Answer: The cause of the sheep's condition is a parasitic infestation affecting the nervous system

### Input:
{meta_knowledge_from_sampling}

Now start generating one question based on the given input.
```

Table 14: Prompt Design for Case Reasoning Data Generation