Towards Omni-RAG: Comprehensive Retrieval-Augmented Generation for Large Language Models in Medical Applications

Zhe Chen[1,3], **Yusheng Liao**[1,3], **Shuyang Jiang**[2,3], **Pingjie Wang**[1,3]
Yiqiu Guo[2,3], **Yanfeng Wang**[1,3], **Yu Wang**[1,3⊠]

[1]Shanghai Jiao Tong University   [2]Fudan University
[3]Shanghai Artificial Intelligence Laboratory

{chenzhe2018,liao20160907,pingjiewang,wangyanfeng622,yuwangsjtu}@sjtu.edu.cn
{shuyangjiang23,yqguo22}@m.fudan.edu.cn

Abstract

Large language models hold promise for addressing medical challenges, such as medical diagnosis reasoning, research knowledge acquisition, clinical decision-making, and consumer health inquiry support. However, they often generate hallucinations due to limited medical knowledge. Incorporating external knowledge is therefore critical, which necessitates multi-source knowledge acquisition. We address this challenge by framing it as a source planning problem, which is to formulate context-appropriate queries tailored to the attributes of diverse sources. Existing approaches either overlook source planning or fail to achieve it effectively due to misalignment between the model's expectation of the sources and their actual content. To bridge this gap, we present MedOmniKB, a repository comprising multigenre and multi-structured medical knowledge sources. Leveraging these sources, we propose the Source Planning Optimisation method, which enhances multi-source utilisation. Our approach involves enabling an expert model to explore and evaluate potential plans while training a smaller model to learn source alignment. Experimental results demonstrate that our method substantially improves multi-source planning performance, enabling the optimised small model to achieve state-of-the-art results in leveraging diverse medical knowledge sources[1].

1 Introduction

Large language models (LLMs) have demonstrated impressive language capabilities (Zhu et al., 2023; Zhou et al., 2023; Chen et al., 2023b; Singhal et al., 2023; Nori et al., 2023; Zhu et al., 2025; Chen et al., 2024b; Saab et al., 2024). However, these models can produce factually incorrect responses—often termed "hallucinations"—when

---

⊠: Corresponding author.
[1]Project website: https://github.com/Jack-ZC8/Omni-RAG-Medical
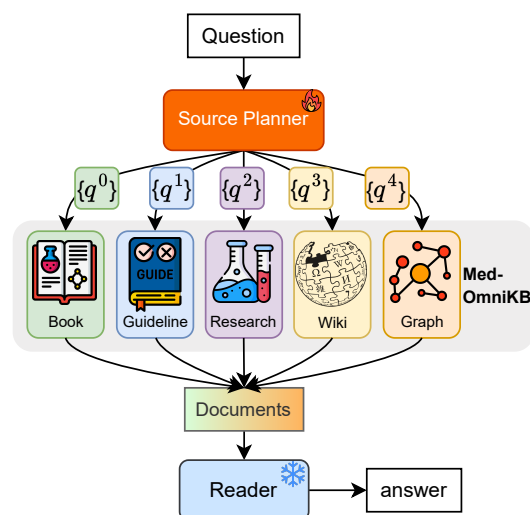

Figure 1: Diagram of source planning in medical scenarios.

their internal knowledge is insufficient (Ji et al., 2023; Jiang et al., 2025). Such inaccuracies pose significant challenges for medical applications, where correctness and trustworthiness are paramount. To mitigate this issue, recent work has explored medical Retrieval-Augmented Generation (RAG) techniques (Xiong et al., 2024a; Yang et al., 2024; Jeong et al., 2024; Xiong et al., 2024b; Xu et al., 2024b), which enhance the accuracy and transparency of model responses (Zhu et al., 2024).

These challenges span various domains, including medical reasoning for diagnoses and treatments, knowledge acquisition from advanced research, clinical decision-making using patient data, and comprehensive responses to health-related inquiries. Addressing these complex healthcare questions requires the integration of diverse knowledge sources (Xiong et al., 2024a; Xu et al., 2024a; Corbeil, 2024). Existing methods typically treat all sources uniformly, using the original question to retrieve without tailoring the search strategy to different sources (Xiong et al., 2024a; Xu et al., 2024a; Neupane et al., 2024; Jeong et al., 2024). Although

15285

subsequent approaches have introduced prompting strategies to guide LLMs in leveraging retrieval sources (Ma et al., 2023; Jiang et al., 2024; Li et al., 2024b), these still fail to construct proper queries for each source. The restricted information presented in the source description prompts causes the misalignment between what the model expects from the sources and what the sources actually contain. Similarly, recent reflection methods (Shinn et al., 2023; Hu et al., 2024; Liao et al., 2024b; Zhao et al., 2024), which iteratively retrieve and self-improve queries, also struggle due to the limited self-reflective capacity when facing multiple sources. In essence, current works cannot construct appropriate queries for each knowledge source based on its unique attributes, which is referred to as the "source planning" problem in our work (see Figure 1).

A key obstacle in studying the problem is the lack of sufficiently broad and diverse medical knowledge bases. To address this, we introduce MedOmniKB, a more comprehensive and varied knowledge repository than previously available resources (Xiong et al., 2024a; Corbeil, 2024; Xu et al., 2024a; Lin et al., 2024b). Five representative sources—"Book," "Guideline," "Research," "Wiki," and "Graph"—offer both depth and breadth of information. Its diversity in categories allows for a more meaningful exploration of source planning.

Leveraging MedOmniKB, we propose a Source Planning Optimisation (SPO), a novel paradigm for knowledge integration. Our approach proceeds as follows: First, we use an expert LLM to perform planning exploration for each question, generating candidate sub-plans. Next, we evaluate whether the documents retrieved by these sub-plans support the correct answer, thereby creating positive and negative plan sets. Finally, we employ a smaller language model to undergo supervised fine-tuning (SFT) on the positive plans and then apply Direct Preference Optimisation (DPO) (Rafailov et al., 2023) to further align the model with the knowledge sources. Extensive experiments show that SPO substantially boosts multi-source planning capability compared to existing techniques. Notably, our optimised small model outperforms substantially larger models (with 10 times the parameters) in retrieval planning. Furthermore, SPO demonstrates robustness under various training conditions and achieves a high utilisation rate of training data.

In summary, our contributions are three-fold:

- We identify the challenge of multi-source plan-

ning in medicine and introduce MedOmniKB as a foundation for research in this area. It addresses the critical gap in richly diverse, large-scale medical knowledge resources.
- Based on MedOmniKB, we propose the SPO approach, a novel paradigm for knowledge integration, empowering language models to adapt retrieval strategies for diverse sources.
- Extensive experiments confirm that the optimised model achieves more efficient source planning than existing methods, offering insights into multi-source retrieval strategies and expanding the applications of large language models in the medical field.

2 Related Work

2.1 Medical Retrieval-augmented Generation

RAG has been successfully applied to a broad spectrum of medical tasks, including clinical decision-making (Shi et al., 2023; Thompson et al., 2023), clinical prediction (Ye et al., 2021; Naik et al., 2022; Xu et al., 2024a), and medical question-answering (QA) (Xiong et al., 2024a; Jeong et al., 2024; Wang et al., 2024d; Li et al., 2024b). Among these applications, medical QA presents significant challenges due to its demand for extensive and precise knowledge integration. While existing studies have developed multi-content retrieval engines for medical information (Xiong et al., 2024a; Corbeil, 2024; Xu et al., 2024a), our approach offers a distinct advantage by utilizing a substantially larger knowledge base that incorporates structured knowledge graphs. As illustrated in Table 1, our knowledge base not only surpasses others in size but also enhances data organization and retrieval accuracy through structured representations.

2.2 Query Construction

Constructing queries is crucial for retrieval augmented generation. Some works directly use large language models to enhance the query (Ma et al., 2023; Wang et al., 2023b, 2024d; Wu et al., 2024b; Wang et al., 2024a; Chen et al., 2024a), and others rely on multiple retrievals and reflection to improve the query quality (Shinn et al., 2023; Hu et al., 2024; Zhao et al., 2024; Liao et al., 2024b). They all struggle to construct proper queries because they either lack the perception of sources or have only limited self-reflective capacity.

There are also studies focused on constructing query-related training data. For example, Ma et al.

Database	#Book	#Guideline	#Research	#Wiki	#Graph
MedCorp (2024a)	9.6k	-	23.9M	6.5M	✗
ClinicalCorp (2024)	9.4k	46.1k	150.4k	-	✗
Self-BioRAG (2024)	18	35.7k	37.5M	-	✗
RAM-EHR (2024a)	-	-	230k	150k	✔
BioKGBench (2024b)	-	-	5.7k	-	✔
MedOmniKB	27.7k	45.7k	25.3M	6.4M	✔

Table 1: Comparison of MedOmniKB with existing medical retrieval knowledge bases in terms of #Docs. "-" denotes the type doesn't exist.

Source	#Docs	#Chunks	#Words/Chunk
Book	27.7k	13.1M	150.1
Guideline	45.7k	647.7k	106.7
Research	25.3M	48.0M	128.7
Wiki	6.4M	29.7M	112.1
	#Concepts	#Definitions	#Relations
Graph	1.7M	317.9k	2.9M

Table 2: Statistics of MedOmniKB.

(2023) evaluate query value based on the performance of the downstream task, which may be influenced by the intrinsic knowledge of LLMs. Mao et al. (2024) employ rerank scores directly, which provide a fast approach but are primarily limited to relevance ranking. Chan et al. (2024) generate three types of enhanced queries using ChatGPT as training data; however, their work does not explicitly assess the quality of these enhanced queries. Yoon et al. (2024); Wang et al. (2024e) incorporate retriever feedback, which can be effective but often relies on gold document annotations, which may not always be available in general scenarios. In contrast, we adopt the LLM-as-a-judge paradigm (Li et al., 2024a) to obtain high-quality training data through LLM-based evaluation. Additionally, while Wang et al. (2023a, 2024b) focus primarily on source selection, they place less emphasis on query reformulation. This highlights the need for further exploration of effective retrieval source planning.

3 MedOmniKB

This section introduces MedOmniKB, a comprehensive multigenre, multistructured medical knowledge base. We investigate the knowledge sources required for existing medical knowledge-intensive problems and correspondingly organise a knowledge base containing five sources: "Book", "Guideline", "Research", "Wiki" and "Graph". The richness and diversity of categories establish the basis for our study of medical source planning. The knowledge base is intended to be used only for research purposes. We did not anonymise them, as they are public medical corpora.

3.1 Unstructured Source

Book. Medical textbooks encompass foundational medical knowledge, which is crucial for understanding the field of medicine. Following Jin et al. (2020); Wu et al. (2024a), 18,182 PDFs are gathered from online libraries and reputable pub-

lishers, whose categories cover medicine, surgery, imaging, etc. Their details are in Appendix D.1. These PDFs were de-duplicated and filtered (URLs, references, citations, etc.) to get documents. We also include StatPearls and Textbooks proposed in MedRAG (Xiong et al., 2024a).

Guideline. Clinical practice guidelines help healthcare practitioners and patients make decisions about diagnosis and treatment. We reuse the guideline data from (Chen et al., 2023a), and for the non-redistributable parts, we crawl webpages using the provided scripts. In the end, we get 45,679 articles from 13 guideline sources.

Research. Research articles offer insights and findings from cutting-edge research, thereby providing a solid theoretical basis for clinical practice and public health decision-making. We download the 2024 PubMed baseline[2], which is a complete snapshot of PubMed data and then extract the valid titles and abstracts.

Wiki. Wikipedia acts as a valuable resource by providing knowledge in general domains. We obtain the processed English data from Huggingface[3].

For retrieval, we process these texts into chunks of no more than 1000 characters, followed by encoding them as vectors using the MedCPT-article-encoder model (Jin et al., 2023). We adopt the Qdrant library[4] for fast dense vector searching.

3.2 Structured Source

Graph. Structured graphs provide clear definitions of concepts and illustrate relationships between them, potentially improving evidence-based decision-making in healthcare. First, the Unified Medical Language System (UMLS) (Bodenreider, 2004) Metathesaurus Full Subset is downloaded

---
[2]https://ftp.ncbi.nlm.nih.gov/pubmed/baseline
[3]https://huggingface.co/datasets/wikimedia/wikipedia
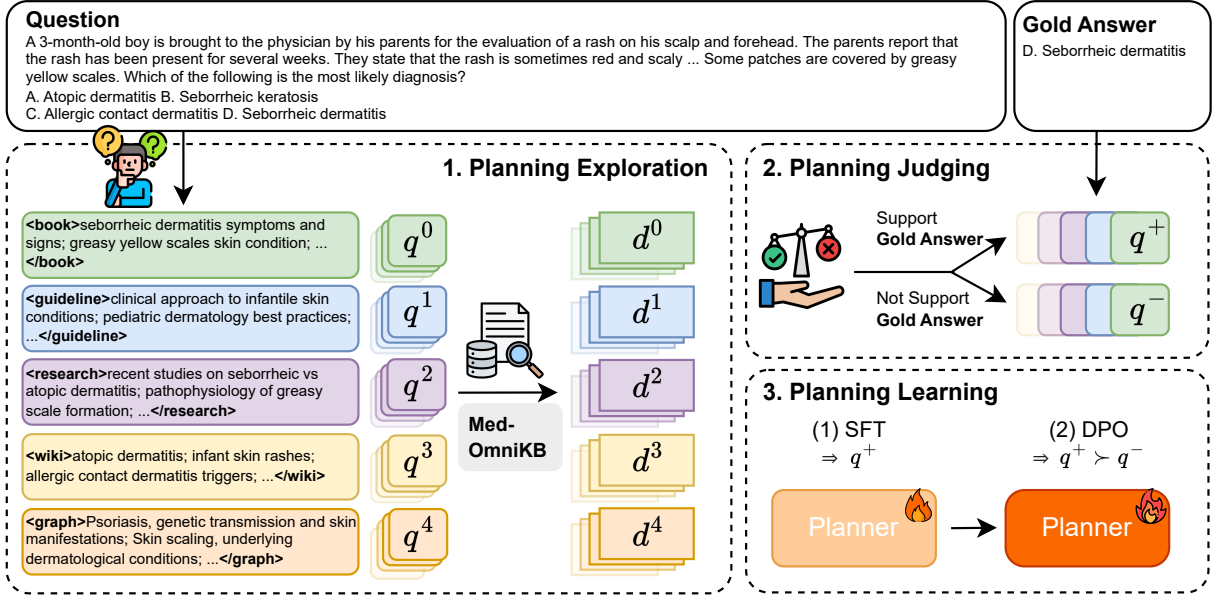[4]https://qdrant.tech

Figure 2: Framework of our proposed SPO approach. First, an expert LLM performs plan exploration for each source, generating multiple queries. Then the expert LLM determines whether the documents retrieved by each query support the gold answer or not. The final positive and negative sets are utilised for planning learning.

and cleaned. Then, the description, indications, pharmacodynamics, absorption, and drug interactions of drugs stored in DrugBank (Wishart et al., 2008) are added to the definition of the corresponding nodes. The combined definitions and relationships are stored in SQLite[5], avoiding the network and latency issues of the online UMLS API.

For a given concept, we will obtain its definition and all its one-hop relationships. Next, reranking will be applied to filter through the many relationships (possibly thousands) following Yang et al. (2023, 2024).

4 Method

4.1 Problem Formulation

We provide a general formulation of our medical source planning problem. For a medical question x, multiple knowledge sources $K = \{K^i \mid i = 1, 2, \ldots, N_K\}$ will be retrieved to supply medical knowledge, where K^i denotes the i_{th} source and N_K denotes the number of sources. Specifically, a planner model $\mathcal{M}_\theta$ will construct the source plan $P = \{(i, j, q_j^i) \mid i = 1, 2, ..., N_K, \ j = 1, 2, ..., N_q^i\}$ for the question x, where q_j^i denotes the j_{th} query in the i_{th} source, and N_q^i represents the number of queries for the i_{th} source. We then obtain the retrieved document $D = \{d_j^i \mid i = 1, 2, ..., N_K, \ j = 1, 2, ..., N_q^i\}$, where d_j^i repre-

sents the top-k documents retrieved by the query q_j^i. Then a language model Reader will read D and answer question x, resulting in an answer y. Our task is to construct the optimal P so that D includes as many documents supporting the gold answer as possible. Additionally, the number of queries per source N_q^i will be limited to fewer than 4 due to the context length restriction of the Reader.

4.2 Source Planning Optimisation

In this section, we introduce our Source Planning Optimisation (SPO) method to empower the planner $\mathcal{M}_\theta$ to retrieve valuable information from multiple healthcare knowledge sources efficiently. Our method comprises three steps: Planning Exploration (Section 4.2.1), Planning Judging (Section 4.2.2) and Planning Learning (Section 4.2.3).

4.2.1 Planning Exploration

We begin with planning exploration to identify potential multi-source planning strategies. This process is performed on a training set consisting of comprehensive, knowledge-intensive medical problems. In our approach, we prompt an expert LLM, Qwen2.5-72B-Instruct-AWQ (Qwen Team, 2024), to generate multiple queries for each source. The exploration prompt is guided by two principles: diversity within a single source and alignment with the characteristics of different sources[6]. The diver-

---

[5] https://www.sqlite.org

[6] The exploration prompt is shown in Prompt E.3.

sity principle encourages a wide range of queries from the same source, promoting varied perspectives and minimizing redundancy. The alignment principle ensures that queries are tailored to fit the unique properties of each source. Experimental results on query diversity are discussed in Appendix A.1, and the number of queries per source is fixed at six.

Subsequently, we use the queries to retrieve their corresponding sources and gather the top-k documents for each query. The superscripts "+" and "−" represent positive and negative ones, respectively.

4.2.2 Planning Judging

Inspired by the recent emergence of LLM-as-a-judge (Li et al., 2024a), we prompt the superior LLM Qwen2.5-72B-Instruct-AWQ to judge whether the documents retrieved by the query support the gold answer[7]. Each q_j^i will be categorized as "positive" or "negative", which can be formulated as:

$$q_j^i = \begin{cases} q_j^{i,+} & d_j^i \text{ support gold answer,} \\ q_j^{i,-} & \text{else,} \end{cases} \quad (1)$$

where d_j^i represents the documents retrieved by the query q_j^i.

4.2.3 Planning Learning

Based on the judgements, we employ $\mathcal{M}_\theta$ to perform supervised fine-tuning (SFT) first, followed by direct preference optimisation (DPO) to further align with the multi-aspect knowledge base.

Supervised Fine-tuning. For the case where more than 3 positive queries exist in a single source, we randomly select 3 of them. For the case where no positive queries exist for a single source, we leave this source empty, meaning it is not used. For the case where no positive queries exist in all sources, we filter the sample out, meaning that none of the sources can provide valuable information. To assess knowledge from all sources, up to 3 positive queries q^+ on each source are ensembled to construct the positive plan P^+ for each question. We perform standard SFT, whose training objective is formulated as:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(x,P^+)\sim\mathcal{D}^+} \log \mathcal{M}_\theta \left(P^+ \mid x \right) \quad (2)$$

---
[7]The judging prompt is shown in Prompt E.4.

Direct Perference Optimisation The planner $\mathcal{M}_\theta$, which builds upon the SFT, is further aligned to multiple source knowledge sources. To achieve this, we collect negative plans for each question in the same manner as the positive plans. The positive and negative plans are then paired for each question to construct the dataset $\mathcal{D}^\pm$. Importantly, we only retain samples where both positive and negative plans are available to ensure balanced training. Subsequently, we employ DPO learning (Rafailov et al., 2023), a method shown to provide stable and effective alignment. The training objective is expressed as:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(x,P^+,P^-)\sim\mathcal{D}^\pm} \log \sigma(r_\theta(x, P^+) \\ - r_\theta(x, P^-)), \quad (3)$$

where

$$r_\theta(x, P) = \beta \log \frac{\mathcal{M}_\theta(P \mid x)}{\mathcal{M}_{\text{SFT}}(P \mid x)}. \quad (4)$$

We expect that this method will enable the planner to achieve precise alignment with complex knowledge sources, thereby facilitating effective multi-source planning.

5 Experiments

5.1 Experimental Settings

5.1.1 Datasets

Dataset	Format	# Train	# Dev	# Test
MedQA		6000	1200	1273
MedMCQA		6000	1200	1200
MMLU-Med		-	-	1089
PubMedQA	Closed-ended	417	83	500
BioASQ	QA	584	116	782
SEER		2113	422	1000
DDXPlus		5000	1000	1000
MIMIC-IV-ED		2917	583	1000
LiveQA		343	68	104
MedicationQA	Open-ended	450	90	150
ExpertQA-Biomed	Generation	375	75	150
Total		24199	4837	8248

Table 3: The statistics of datasets in our experiments. The training and development sets are combined to expect universal planning capabilities.

We conduct our experiments on 11 datasets, which are well-defined and widely adopted for testing LLMs, as shown in Table 3. These tasks represent the core focus of the medical application (detailed in Table 17). Reasoning QA datasets include MedQA (Jin et al., 2020), MedMCQA (Pal et al.,

15289

Planner	Method	Reasoning QA			Research QA		Clinical QA			Average
		MedQA	MedMCQA	MMLU.	PubMedQA	BioASQ	SEER	DDXPlus	MIMIC-IV.	
					Reader: Frozen Qwen2.5-7B					
-	No Retrieval	60.80	56.17	76.95	34.60	74.81	51.00	42.80	58.50	56.95
	Original Question	62.45	63.25	80.90	47.00	89.00	58.40	42.80	57.90	62.71
	Query2Doc	62.92	66.42	80.26	46.40	88.24	<u>58.80</u>	42.40	56.90	62.79
Frozen Qwen2.5-72B	Prompting	72.11	65.33	81.73	53.80	89.64	57.10	48.70	62.00	66.30
	Reflexion	<u>73.13</u>	66.00	79.06	52.60	89.64	57.90	49.40	62.60	66.29
	SeRTS	70.70	<u>66.83</u>	<u>82.55</u>	<u>55.60</u>	**90.03**	57.10	<u>51.20</u>	62.50	<u>67.06</u>
Trained Qwen2.5-7B	Trainable Planning	72.03	66.42	82.19	54.80	<u>89.90</u>	57.20	46.40	60.30	66.16
	RaFe Planning	70.86	66.50	78.70	53.40	89.77	55.20	50.30	<u>63.70</u>	66.05
	SPO Planning	**76.98**	**71.08**	**85.49**	**60.20**	89.77	**61.90**	**52.40**	**69.60**	**70.93**
					Reader: Frozen Llama3.1-8B					
-	No Retrieval	65.99	59.50	76.58	56.20	81.97	57.00	38.80	58.60	61.83
	Original Question	60.57	57.50	72.18	74.20	87.47	57.60	39.00	58.10	63.33
	Query2Doc	61.04	59.92	72.91	<u>74.80</u>	87.21	57.60	39.20	57.70	63.80
Frozen Qwen2.5-72B	Prompting	71.17	62.08	75.94	71.40	89.00	57.50	41.10	58.60	65.85
	Reflexion	<u>72.82</u>	63.17	75.48	71.80	<u>89.51</u>	56.80	41.40	59.70	66.34
	SeRTS	71.88	63.25	<u>77.13</u>	71.60	<u>89.51</u>	57.00	42.90	60.10	<u>66.67</u>
Trained Qwen2.5-7B	Trainable Planning	71.64	62.33	75.76	72.20	89.00	58.30	<u>43.60</u>	<u>60.20</u>	66.63
	RaFe Planning	69.76	<u>63.67</u>	74.75	72.20	89.00	<u>58.70</u>	42.40	59.40	66.24
	SPO Planning	**77.45**	**69.25**	**78.97**	**75.60**	**89.64**	**60.70**	**45.70**	**64.10**	**70.18**
					Reader: Frozen Mistral0.3-7B					
-	No Retrieval	49.18	45.58	63.54	47.40	73.27	49.40	23.50	58.20	51.26
	Original Question	53.18	58.00	68.96	**67.00**	**88.87**	51.10	30.60	57.00	59.34
	Query2Doc	52.40	<u>59.67</u>	70.71	<u>66.60</u>	86.83	52.90	30.50	57.00	59.58
Frozen Qwen2.5-72B	Prompting	64.26	57.75	72.73	59.80	<u>88.75</u>	55.10	41.10	58.60	62.26
	Reflexion	64.26	57.50	71.44	58.40	87.21	<u>56.50</u>	41.70	58.50	61.94
	SeRTS	<u>65.59</u>	58.83	<u>73.00</u>	58.60	87.34	56.40	<u>43.50</u>	58.90	<u>62.77</u>
Trained Qwen2.5-7B	Trainable Planning	62.45	55.75	71.63	56.80	87.08	53.30	41.70	58.40	60.89
	RaFe Planning	61.67	55.25	71.99	56.20	82.35	56.20	41.20	<u>60.40</u>	60.66
	SPO Planning	**71.09**	**64.25**	**75.02**	65.20	87.60	**59.00**	**45.20**	**64.80**	**66.52**

Table 4: Comparison of our method with baselines. **Bold** represents the best result, while <u>underlining</u> represents the second-best result. All models are in the Instruct version. AWQ is applied to the models of 72B size.

Method	Relevance	Completeness	Proficiency	Interpretation
No Retrieval	**3.83**	3.41	3.02	-
	(4.1/3.7/3.6)	(3.3/3.6/3.4)	(2.8/3.6/2.7)	-
Original Question	3.79	3.55	3.48	3.29
	(4.1/3.5/3.7)	(3.6/3.5/3.6)	(3.5/3.3/3.6)	(3.0/3.2/3.7)
SeRTS	3.80	3.49	3.30	3.19
	(3.9/3.9/3.6)	(3.3/3.8/3.3)	(3.6/3.6/2.7)	(3.4/3.0/3.1)
SPO Planning	3.79	**4.18**	**4.08**	**3.97**
	(4.1/3.7/3.6)	(4.3/4.0/4.3)	(4.2/4.0/4.1)	(3.8/4.1/4.0)

Table 5: Expert evaluation results of open-ended generation datasets. The reader is frozen Qwen2.5-7B. Each cell reports the average value (LiveQA/MedicationQA/ExpertQA-Biomed).

2022), and MMLU-Med (Hendrycks et al., 2021). Research QA datasets include PubMedQA (Jin et al., 2019) and BioASQ (Tsatsaronis et al., 2015). Clinical QA datasets include SEER (Dubey et al., 2023), DDXPlus (Tchango et al., 2022) and MIMIC-IV-ED (Xie et al., 2022). Long-form Answering datasets include LiveQA (Abacha et al., 2017), MedicationQA (Abacha et al., 2019) and ExpertQA-Biomed (Malaviya et al., 2024). More details of the datasets can be found in Appendix D.2. The accuracy is adopted as the metric for closed-ended QA datasets, while expert evalu-ation is used for open-ended generation datasets. The details of expert evaluation are shown in Appendix A.4.

5.1.2 Retrieval Procedure

The proposed MedOmniKB is adopted as our retrieval knowledge base. For the unstructured sources, the format of the queries is "query0; query1; ...". We retrieve the top-20 documents and then rerank the top-10 documents using MedCPT models (Jin et al., 2023). For the structured source, the format of the queries is "term0, query0; term0, query1; ...". We search for the definition of each "term" and all its reranked top-10 relations. The latter half of the search results is excluded during the planning judging phase to enhance efficiency.

5.2 Baselines

Given the limited research on efficient multi-source planning for medical information retrieval, we replicate single-source querying approaches as baselines. Unsupervised methods include No Retrieval, Original Question (Xiong et al., 2024a), Query2Doc (Wang et al., 2023b),

Sources	MedQA	PubMedQA	SEER
All	76.98	60.20	61.90
- Book	70.38 (-8.57%)	52.60 (-12.62%)	56.80 (-8.24%)
- Guideline	72.35 (-6.01%)	56.40 (-6.31%)	52.10 (-15.83%)
- Research	71.72 (-6.83%)	35.40 (-41.20%)	51.20 (-17.29%)
- Wiki	72.35 (-6.01%)	58.20 (-3.32%)	55.50 (-10.34%)
- Graph	71.48 (-7.14%)	58.20 (-3.32%)	56.80 (-8.24%)
No	60.80 (-21.02%)	34.60 (-42.52%)	51.00 (-17.61%)

Table 6: Accuracy and relative drop ratio of reader when dropping documents from each source.

Figure 3: Average accuracy of different sizes of readers paired with different planners.

Prompting (Ma et al., 2023), Reflexion (Shinn et al., 2023), and SeRTS (Hu et al., 2024), all of which are adapted to support multi-source planning. Supervised methods consist of Trainable Planning (Ma et al., 2023) and RaFe Planning (Mao et al., 2024), implemented with modifications by adjusting the judgment signal, specifically using QA accuracy (Rouge-L (Lin, 2004) for open-ended generation) and rerank scores for query evaluation, respectively. Additional details on the baseline methods can be found in Appendix D.3.

5.3 Main Results

To evaluate the effectiveness of our method, we incorporate three frozen readers with varying intrinsic knowledge: Qwen2.5-7B (Qwen Team, 2024), Llama3.1-8B (Dubey et al., 2024), and Mistral0.3-7B (Jiang et al., 2023). The use of these pretrained models is consistent with their intended purposes. The main results are summarized in Table 4 and Table 5, leading to the following key observations:

1. Effectiveness of Planning Methods: Planning approaches (baselines excluding Original Question and Query2Doc) consistently outperform direct retrieval methods that rely on the original question or pseudo-document queries across most test cases. This underscores the importance of multi-source planning, which enables language models to retrieve information dynamically from diverse sources. However, for PubMedQA and BioASQ, non-planning approaches yield competitive results. This can be attributed to the strong alignment between these datasets and the "Research" source, reducing the need for planning.

2. Superiority of SPO Training Signals: The SPO method demonstrates superior training signals compared to existing trainable approaches. SPO avoids the risk of intrinsic knowledge overshadowing retrieved knowledge (Ma et al., 2023). Additionally, it provides more accurate judgments than those proposed in Mao et al. (2024), further vali-
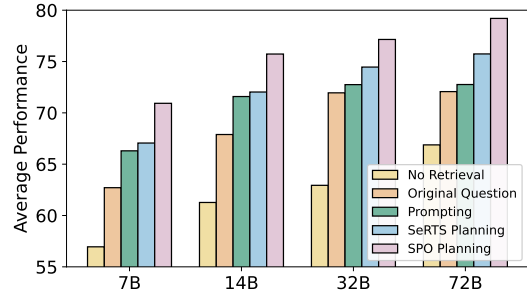
dating the effectiveness of our method.

3. Performance Across Readers: When paired with each reader, SPO planning achieves the best performance in most configurations, outperforming models with up to 10 times the number of parameters. Furthermore, SPO can maximally enhance the completeness, proficiency, and interpretation of LLM medical responses. These improvements highlight SPO planning's ability to provide readers with abundant and relevant knowledge.

5.4 Enhancement for Different Sizes of Readers

We also use models of different sizes as readers. We use frozen 7B, 14B, 32B, and 72B Qwen2.5-Instruct models as readers and show the average accuracy on all closed-ended datasets as shown in Figure 3. AWQ is applied to models of 32B and 72B sizes. In general, the performances of the readers in all cases rise with increasing model size, which is attributed to the increasing intrinsic knowledge. Significantly, it can be found that the SPO planner brings the most enhancements among all planners, even as the readers' intrinsic knowledge increases.

5.5 Effectiveness of Different Sources

To investigate the effects of each retrieval source, we systematically exclude documents from each source which is retrieved using the SPO planner. The results are shown in Table 6. We utilise the frozen Qwen2.5-7B-Instruct model as the reader. Overall, every retrieval source enhances the ability to answer medical questions to varying degrees, with the "Book", "Guideline", and "Research" sources providing the most significant improvements. Regarding the impact of sources on different types of QA, the "Research" source notably impacts performance on the PubMedQA dataset. It aligns with the "Research" style inherent to the
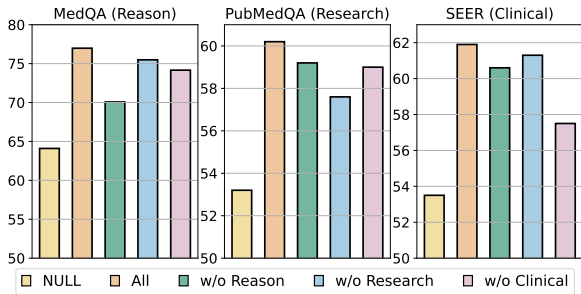
Figure 4: Accuracy of the reader paired with SPO planner trained using different categories of data.

Training Strategy	MedQA	PubMedQA	SEER
Frozen	64.10	53.20	53.50
+ SFT	74.08	59.20	58.50
+ SFT + DPO	**76.98**	**60.20**	**61.90**
+ DPO	67.48	55.80	54.30

Table 7: Accuracy of the reader paired with SPO planner trained using different strategies.

dataset. The "Guideline" and "Research" sources have a significant impact on SEER, which can be attributed to its characteristics as a treatment planning dataset. Moreover, the visualization of source planning statistics is shown in Appendix B.1.

5.6 Effectiveness of Training Stages

To examine the impact of the two training phases, SFT and DPO, we present the results of the planner model trained under different configurations in Table 7. Frozen Qwen2.5-7B-Instruct is used as the reader. It is important to note that the Frozen model referenced here is 7B, whereas the Frozen model in Table 4 is 72B. The results demonstrate that SFT significantly enhances the model's multi-source planning capability, providing a strong foundational ability. The subsequent DPO phase further refines and strengthens this capability. However, applying DPO directly to the planner yields only limited improvements, underscoring the critical role of the SFT phase in establishing the model's initial planning proficiency.

5.7 Adaptability to Out-of-Distribution Data

We classify the out-of-distribution (OOD) task into three categories according to Table 17, where we systematically exclude the training data for individual categories and assess the planner's OOD performance. We use the frozen Qwen2.5-7B-Instruct as the reader. Figure 4 compares the performance of models without training, trained with the full
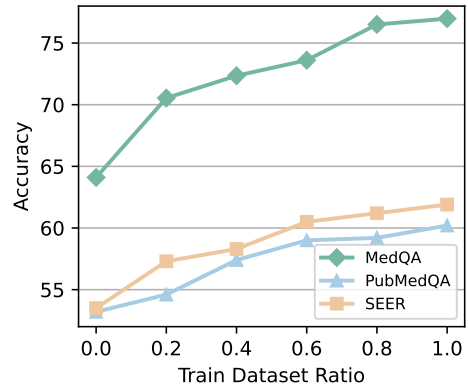

Figure 5: Accuracy of the reader paired with SPO planner trained using different ratios of data.

Training Strategy	MedQA*	PubMedQA*	SEER*
No Retrieval	60.80	34.60	51.00
Frozen	62.77	47.00	52.50
+ SPO	69.21	58.20	56.80
+ SPO (OOD corpus)	65.44	50.60	53.20

Table 8: Accuracy of the reader paired with SPO planner trained using different strategies. The asterisk (*) indicates that only a single retrieval source is available during the testing phase. Specifically, only the Book source is available on MedQA, only the Research source is available on PubMedQA, and only the Guideline source is available on SEER.

dataset, and trained with data from each type of data excluded. Firstly, the SPO planner outperforms the untrained planner across the test sets of all three domains under the out-of-type setting (for example, MedQA in the "w/o Reason" case). This demonstrates that the SPO method can learn a generalized source planning ability, regardless of the category of training data. Furthermore, it can be observed that when training data of a different type from the current test set is dropped (for example, MedQA in the "w/o Research" case), the model's performance slightly declines. This further confirms that all types of data contribute to learning a universal planning ability.

5.8 Adaptability to Out-of-Distribution Retrieval Source

To explore the SPO's adaptability to an unknown retrieval corpus, we designed a new baseline, SPO (OOD corpus). We select "Book" as the unknown corpus for MedQA, "Research" for PubMedQA, and "Guideline" for SEER. During training, only the unknown source is excluded in the prompt and label (a total of 4 sources); however, during testing, only this source is included in the prompt (a total

of 1 source). All methods' test prompts remain the same. Frozen Qwen2.5-7B is adopted as the reader.

Table 8 shows that the SPO (OOD corpus) planner can retrieve more closely aligned with the characteristics of unknown sources compared to the frozen planner. The SPO method enhances the perception of the source, even though the source is not included in the training set. It is reasonable, as all sources share semantic similarities in the medical field. The adaptability of the SPO method to OOD corpora broadens its potential applications.

5.9 Robustness to Training Data Quantity

We further examine the impact of data quantity by sampling different proportions of the training dataset and applying the same experimental setup as in the main study. We use the frozen Qwen2.5-7B-Instruct as the reader. As shown in Figure 5, planning performance improves consistently with increasing data volume but begins to plateau at higher levels. These results indicate that the SPO method significantly enhances planning capability even with limited data, while larger datasets provide diminishing yet notable additional benefits.

6 Conclusion

This study tackles the challenge of multi-source planning in healthcare by introducing the MedOmniKB knowledge base and proposing the Source Planning Optimisation method. The MedOmniKB combines diverse medical information sources with high relevance. The Source Planning Optimisation method significantly enhances language models' ability to retrieve from various knowledge sources. Extensive experiments show our optimised small model has outperformed the larger one in planning efficiency. In this work, we underscore the importance of effective source planning. Our model effectively integrates multi-source knowledge, advancing the application of language models in the healthcare field.

Limitations

There are some limitations in our work. First, our constructed MedOmniKB may not encompass all medical knowledge resources. However, the five sources we select are not only representative but also essential for various applications requiring medical knowledge. Other medical knowledge not covered by the five sources still needs to be identified and incorporated into MedOmniKB.

Moreover, the exploration and judging steps in our SPO approach incur a relatively large retrieval and inference cost. The retrieval cost arises from the large and diverse retrieval repository, while the inference cost is due to the separate evaluation for each document set. In our opinion, the accurate judgement of each query is essential for the model to learn source planning. We alleviate the cost problems by optimising the retrieval framework based on Qdrant and multi-process parallelism. Comparison of cost with existing supervised methods and cost analysis in multi-process scenarios are presented in Appendix B.3.

Lastly, there are limitations in our evaluation datasets and methods. Although we have included as many categories of medical datasets as possible, there may still be unexplored specific medical scenarios. Additionally, while we have employed evaluation methods such as accuracy for question-answering, and expert assessments for long-text generation, further evaluation is still needed. This should include assessing user satisfaction, treatment outcomes, and other factors in real medical scenarios.

Ethical Consideration

The deployment of large language models (LLMs) in healthcare raises significant ethical challenges that warrant careful consideration. Ensuring accuracy is paramount, as inaccuracies may adversely affect patient outcomes. Our Source Planning Optimisation (SPO) method addresses this concern by enhancing the retrieval of information from credible sources. It can help models generate accurate and reliable responses.

Furthermore, transparency in model outputs is markedly improved through access to the retrieved knowledge. By documenting the sources of information utilized in generating responses, healthcare professionals can understand the foundations of the information presented to them. This transparency also enables practitioners to assess the reliability of the model responses.

Acknowledgements

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 *Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA, November 15-17, 2017*, volume 500-324 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

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 - Proceedings of the 17th World Congress on Medical and Health Informatics, Lyon, France, 25-30 August 2019*, volume 264 of *Studies in Health Technology and Informatics*, pages 25–29. IOS Press.

Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, 32(Database-Issue):267–270.

Chi-Min Chan, Chunpu Xu, Ruibin Yuan, Hongyin Luo, Wei Xue, Yike Guo, and Jie Fu. 2024. RQ-RAG: learning to refine queries for retrieval augmented generation. *CoRR*, abs/2404.00610.

Guanhua Chen, Wenhan Yu, and Lei Sha. 2024a. Unlocking multi-view insights in knowledge-dense retrieval-augmented generation. *CoRR*, abs/2404.12879.

Zeming Chen, Alejandro Hernández-Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023a. MEDITRON-70B: scaling medical pretraining for large language models. *CoRR*, abs/2311.16079.

Zhe Chen, Heyang Liu, Wenyi Yu, Guangzhi Sun, Hongcheng Liu, Ji Wu, Chao Zhang, Yu Wang, and Yanfeng Wang. 2024b. M^{3}AV: A multimodal, multi-genre, and multipurpose audio-visual academic lecture dataset. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 9041–9060. Association for Computational Linguistics.

Zhe Chen, Hongcheng Liu, and Yu Wang. 2023b. DialogMCF: Multimodal context flow for audio visual scene-aware dialog. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:753–764.

Jean-Philippe Corbeil. 2024. Iryonlp at MEDIQA-CORR 2024: Tackling the medical error detection & correction task on the shoulders of medical agents. *CoRR*, abs/2404.15488.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Snigdha Dubey, Gaurav Tiwari, Sneha Singh, Saveli Goldberg, and Eugene Pinsky. 2023. Using machine learning for healthcare treatment planning. *Frontiers Artif. Intell.*, 6.

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.

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 *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Minda Hu, Licheng Zong, Hongru Wang, Jingyan Zhou, Jingjing Li, Yichen Gao, Kam-Fai Wong, Yu Li, and Irwin King. 2024. Serts: Self-rewarding tree search for biomedical retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 1321–1335. Association for Computational Linguistics.

Minbyul Jeong, Jiwoong Sohn, Mujeen Sung, and Jaewoo Kang. 2024. Improving medical reasoning through retrieval and self-reflection with

retrieval-augmented large language models. *CoRR*, abs/2401.15269.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Shuyang Jiang, Yusheng Liao, Zhe Chen, Ya Zhang, Yanfeng Wang, and Yu Wang. 2025. MedS^3: Towards medical small language models with self-evolved slow thinking. *CoRR*, abs/2501.12051.

Xinke Jiang, Yuchen Fang, Rihong Qiu, Haoyu Zhang, Yongxin Xu, Hao Chen, Wentao Zhang, Ruizhe Zhang, Yuchen Fang, Xu Chu, Junfeng Zhao, and Yasha Wang. 2024. Tc-rag:turing-complete rag's case study on medical LLM systems. *CoRR*, abs/2408.09199.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *CoRR*, abs/2009.13081.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *CoRR*, abs/1909.06146.

Qiao Jin, Won Kim, Qingyu Chen, Donald C. Comeau, Lana Yeganova, W. John Wilbur, and Zhiyong Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinform.*, 39(10).

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. *CoRR*, abs/2309.06180.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2024a. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv: 2411.16594*.

Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Shafiq Joty, Soujanya Poria, and Lidong Bing. 2024b. Chain-of-knowledge: Grounding large language models via dynamic knowledge adapting over heterogeneous sources. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Yusheng Liao, Shuyang Jiang, Zhe Chen, Yanfeng Wang, and Yu Wang. 2024a. Medcare: Advancing medical llms through decoupling clinical alignment and knowledge aggregation. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 10562–10581. Association for Computational Linguistics.

Yusheng Liao, Shuyang Jiang, Yanfeng Wang, and Yu Wang. 2024b. Reflectool: Towards reflection-aware tool-augmented clinical agents. *CoRR*, abs/2410.17657.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024a. AWQ: activation-aware weight quantization for on-device LLM compression and acceleration. In *Proceedings of the Seventh Annual Conference on Machine Learning and Systems, MLSys 2024, Santa Clara, CA, USA, May 13-16, 2024*. mlsys.org.

Xinna Lin, Siqi Ma, Junjie Shan, Xiaojing Zhang, Shell Xu Hu, Tiannan Guo, Stan Z. Li, and Kaicheng Yu. 2024b. Biokgbench: A knowledge graph checking benchmark of AI agent for biomedical science. *CoRR*, abs/2407.00466.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting for retrieval-augmented large language models. *CoRR*, abs/2305.14283.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2024. Expertqa: Expert-curated questions and attributed answers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 3025–3045. Association for Computational Linguistics.

Shengyu Mao, Yong Jiang, Boli Chen, Xiao Li, Peng Wang, Xinyu Wang, Pengjun Xie, Fei Huang, Huajun Chen, and Ningyu Zhang. 2024. Rafe: Ranking feedback improves query rewriting for RAG. *CoRR*, abs/2405.14431.

Aakanksha Naik, Sravanthi Parasa, Sergey Feldman, Lucy Lu Wang, and Tom Hope. 2022. Literature-augmented clinical outcome prediction. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 438–453. Association for Computational Linguistics.

Subash Neupane, Shaswata Mitra, Sudip Mittal, Noorbakhsh Amiri Golilarz, Shahram Rahimi, and Amin

Amirlatifi. 2024. Medinsight: A multi-source context augmentation framework for generating patient-centric medical responses using large language models. *CoRR*, abs/2403.08607.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of GPT-4 on medical challenge problems. *CoRR*, abs/2303.13375.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering. *CoRR*, abs/2203.14371.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *CoRR*, abs/2305.18290.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, Juanma Zambrano Chaves, Szu-Yeu Hu, Mike Schaekermann, Aishwarya Kamath, Yong Cheng, David G. T. Barrett, Cathy Cheung, Basil Mustafa, Anil Palepu, Daniel McDuff, Le Hou, Tomer Golany, Luyang Liu, Jean-Baptiste Alayrac, Neil Houlsby, Nenad Tomasev, Jan Freyberg, Charles Lau, Jonas Kemp, Jeremy Lai, Shekoofeh Azizi, Kimberly Kanada, SiWai Man, Kavita Kulkarni, Ruoxi Sun, Siamak Shakeri, Luheng He, Benjamin Caine, Albert Webson, Natasha Latysheva, Melvin Johnson, Philip Andrew Mansfield, Jian Lu, Ehud Rivlin, Jesper Anderson, Bradley Green, Renee Wong, Jonathan Krause, Jonathon Shlens, Ewa Dominowska, S. M. Ali Eslami, Katherine Chou, Claire Cui, Oriol Vinyals, Koray Kavukcuoglu, James Manyika, Jeff Dean, Demis Hassabis, Yossi Matias, Dale R. Webster, Joelle K. Barral, Greg Corrado, Christopher Semturs, S. Sara Mahdavi, Juraj Gottweis, Alan Karthikesalingam, and Vivek Natarajan. 2024. Capabilities of gemini models in medicine. *CoRR*, abs/2404.18416.

Wenqi Shi, Yuchen Zhuang, Yuanda Zhu, Henry J. Iwinski, J. Michael Wattenbarger, and May Dongmei Wang. 2023. Retrieval-augmented large language models for adolescent idiopathic scoliosis patients in shared decision-making. In *Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, BCB 2023, Houston, TX, USA, September 3-6, 2023*, pages 14:1–14:10. ACM.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Andrew Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Agüera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle K. Barral, Dale R. Webster, Gregory S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards expert-level medical question answering with large language models. *CoRR*, abs/2305.09617.

Arsène Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. 2022. Ddxplus: A new dataset for automatic medical diagnosis. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Will E. Thompson, David M. Vidmar, Jessica K. De Freitas, John M. Pfeifer, Brandon K. Fornwalt, Ruijun Chen, Gabriel Altay, Kabir Manghnani, Andrew C. Nelsen, Kellie Morland, Martin C. Stumpe, and Riccardo Miotto. 2023. Large language models with retrieval-augmented generation for zero-shot disease phenotyping. *CoRR*, abs/2312.06457.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artières, Axel-Cyrille Ngonga Ngomo, Norman Heino, Éric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. 2015. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinform.*, 16:138:1–138:28.

Haoyu Wang, Ruirui Li, Haoming Jiang, Jinjin Tian, Zhengyang Wang, Chen Luo, Xianfeng Tang, Monica Xiao Cheng, Tuo Zhao, and Jing Gao. 2024a. Blendfilter: Advancing retrieval-augmented large language models via query generation blending and knowledge filtering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 1009–1025. Association for Computational Linguistics.

Hongru Wang, Minda Hu, Yang Deng, Rui Wang, Fei Mi, Weichao Wang, Yasheng Wang, Wai-Chung Kwan, Irwin King, and Kam-Fai Wong. 2023a. Large language models as source planner for personalized knowledge-grounded dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 9556–9569. Association for Computational Linguistics.

Hongru Wang, Wenyu Huang, Yang Deng, Rui Wang, Zezhong Wang, Yufei Wang, Fei Mi, Jeff Z. Pan,

and Kam-Fai Wong. 2024b. Unims-rag: A unified multi-source retrieval-augmented generation for personalized dialogue systems. *CoRR*, abs/2401.13256.

Liang Wang, Nan Yang, and Furu Wei. 2023b. Query2doc: Query expansion with large language models. *CoRR*, abs/2303.07678.

Xidong Wang, Guiming Chen, Dingjie Song, Zhiyi Zhang, Zhihong Chen, Qingying Xiao, Junying Chen, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. 2024c. CMB: A comprehensive medical benchmark in chinese. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 6184–6205. Association for Computational Linguistics.

Yubo Wang, Xueguang Ma, and Wenhu Chen. 2024d. Augmenting black-box llms with medical textbooks for biomedical question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 1754–1770. Association for Computational Linguistics.

Yujing Wang, Hainan Zhang, Liang Pang, Binghui Guo, Hongwei Zheng, and Zhiming Zheng. 2024e. Maferw: Query rewriting with multi-aspect feedbacks for retrieval-augmented large language models. *CoRR*, abs/2408.17072.

David S. Wishart, Craig Knox, Anchi Guo, Dean Cheng, Savita Shrivastava, Dan Tzur, Bijaya Gautam, and Murtaza Hassanali. 2008. Drugbank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.*, 36(Database-Issue):901–906.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024a. Pmc-llama: toward building open-source language models for medicine. *J. Am. Medical Informatics Assoc.*, 31(9):1833–1843.

Chaoyi Wu, Pengcheng Qiu, Jinxin Liu, Hongfei Gu, Na Li, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2025. Towards evaluating and building versatile large language models for medicine. *npj Digit. Medicine*, 8(1).

Ridong Wu, Shuhong Chen, Xiangbiao Su, Yuankai Zhu, Yifei Liao, and Jianming Wu. 2024b. A multi-source retrieval question answering framework based on RAG. *CoRR*, abs/2405.19207.

Feng Xie, Jun Zhou, Jin Wee Lee, Mingrui Tan, Siqi Li, Logasan S/O Rajnthern, Marcel Lucas Chee, Bibhas Chakraborty, An-Kwok Ian Wong, Alon Dagan, Marcus Eng Hock Ong, Fei Gao, and Nan Liu. 2022. Benchmarking emergency department triage prediction models with machine learning and large public electronic health records. In *AMIA 2022, American Medical Informatics Association Annual Symposium, Washington, DC, USA, November 5-9, 2022*. AMIA.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024a. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 6233–6251. Association for Computational Linguistics.

Guangzhi Xiong, Qiao Jin, Xiao Wang, Minjia Zhang, Zhiyong Lu, and Aidong Zhang. 2024b. Improving retrieval-augmented generation in medicine with iterative follow-up questions. *CoRR*, abs/2408.00727.

Ran Xu, Wenqi Shi, Yue Yu, Yuchen Zhuang, Bowen Jin, May D. Wang, Joyce C. Ho, and Carl Yang. 2024a. RAM-EHR: retrieval augmentation meets clinical predictions on electronic health records. *CoRR*, abs/2403.00815.

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. *CoRR*, abs/2404.18443.

Rui Yang, Haoran Liu, Edison Marrese-Taylor, Qingcheng Zeng, Yuhe Ke, Wanxin Li, Lechao Cheng, Qingyu Chen, James Caverlee, Yutaka Matsuo, and Irene Li. 2024. Kg-rank: Enhancing large language models for medical QA with knowledge graphs and ranking techniques. *CoRR*, abs/2403.05881.

Rui Yang, Edison Marrese-Taylor, Yuhe Ke, Lechao Cheng, Qingyu Chen, and Irene Li. 2023. Integrating UMLS knowledge into large language models for medical question answering. *CoRR*, abs/2310.02778.

Muchao Ye, Suhan Cui, Yaqing Wang, Junyu Luo, Cao Xiao, and Fenglong Ma. 2021. Medretriever: Target-driven interpretable health risk prediction via retrieving unstructured medical text. In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, pages 2414–2423. ACM.

Chanwoong Yoon, Gangwoo Kim, Byeongguk Jeon, Sungdong Kim, Yohan Jo, and Jaewoo Kang. 2024. Ask optimal questions: Aligning large language models with retriever's preference in conversational search. *CoRR*, abs/2402.11827.

Caiqi Zhang, Fangyu Liu, Marco Basaldella, and Nigel Collier. 2024. LUQ: Long-text uncertainty quantification for LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5244–5262, Miami, Florida, USA. Association for Computational Linguistics.

Qingfei Zhao, Ruobing Wang, Xin Wang, Daren Zha, and Nan Mu. 2024. Towards multi-source retrieval-augmented generation via synergizing reasoning and preference-driven retrieval. *CoRR*, abs/2411.00689.

Hongjian Zhou, Boyang Gu, Xinyu Zou, Yiru Li, Sam S. Chen, Peilin Zhou, Junling Liu, Yining Hua, Chengfeng Mao, Xian Wu, Zheng Li, and Fenglin Liu. 2023. A survey of large language models in medicine: Progress, application, and challenge. *CoRR*, abs/2311.05112.

Zhiyuan Zhu, Yusheng Liao, Zhe Chen, Yu Wang, and Yunfeng Guan. 2023. Towards optimizing pretrained language model ensemble learning for task-oriented dialogue system. In *Proceedings of The Eleventh Dialog System Technology Challenge*, pages 144–149.

Zhiyuan Zhu, Yusheng Liao, Zhe Chen, Yuhao Wang, Yunfeng Guan, Yanfeng Wang, and Yu Wang. 2025. Evolvebench: A comprehensive benchmark for assessing temporal awareness in llms on evolving knowledge. In *ACL 2025*.

Zhiyuan Zhu, Yusheng Liao, Chenxin Xu, Yunfeng Guan, Yanfeng Wang, and Yu Wang. 2024. RA2FD: Distilling faithfulness into efficient dialogue systems. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12304–12317, Miami, Florida, USA. Association for Computational Linguistics.

A   Additional Experiments

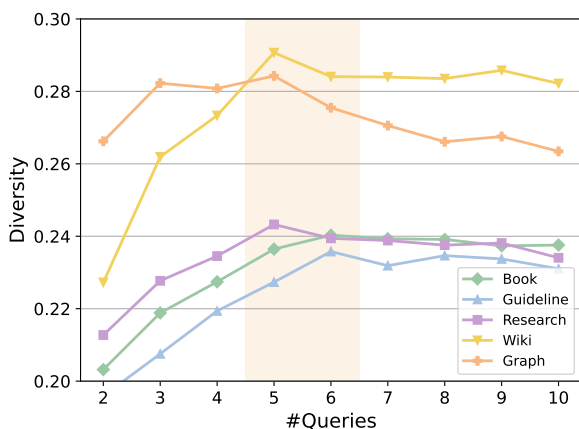## A.1   Optimal Number of Queries in Planning Exploration

Figure 6: The query diversity of varying queries on the development set.

In this section, we aim to identify the optimal number of queries for each source in Planning Exploration (4.2.1). Our objective is to balance the need for diverse queries with the constraints of retrieval and inference costs, which can become significant with excessive querying. To achieve this, we define the optimal number based on the diversity of the source's queries, as our primary goal is to ensure diverse exploration. Specifically,

the diversity of a source's queries q^i is quantified as:

$$D = \frac{1}{\binom{N}{2}} \sum_{j=1}^{N} \sum_{k=j+1}^{N} \text{distance}(q_j^i, q_k^i), \quad (5)$$

where N is the number of $\{q^i\}$ and $\text{distance}(\cdot, \cdot)$ is the cosine distance between encoded query vectors.

Figure 6 illustrates the query diversity across different sources on the development set. In the SPO approach, we set the number of queries per source to 6, as the diversity peaks when the number of queries ranges between 5 and 6.

A.2   Human Evaluation of LLM Judgements

We present a manual evaluation of existing methods for constructing query training data. We randomly sample 800 samples from the training set. Medical researchers are employed to determine whether each document supports the correct answer, yielding 800 tuples of (question + correct answer, document, human judgement). Subsequently, we treat the discrimination results of different methods as hypotheses, and human judgment as references and calculate accuracy, precision, recall and F1 scores.

The human evaluation results are illustrated in Table 9. It is clear that the SPO method yields much superior query quality than the existing methods. As mentioned earlier, Trainable Planning (Ma et al., 2023) evaluates query value based on downstream task performance, potentially influenced by LLMs' intrinsic knowledge. RaFe Planning (Mao et al., 2024) uses rerank scores directly, offering a quick method but mainly focused on relevance ranking. Moreover, our findings indicate that LLMs possess a sufficient capacity to perform the judgment task, even with the smaller Qwen2.5-32B-Instruct model. This illustrates the superiority of the SPO method.

A.3   Assessment of Uncertainty Introduced by Retrieval

Experiments are conducted to assess the uncertainty brought by the SPO method. The uncertainty of Reader responses reflects the ambiguity present in the retrieved information. We reimplement Long-Text Uncertainty Quantification (LUQ) (Zhang et al., 2024) to assess the uncertainty of LLMs. The temperature is set to 0.7, and the number of samples is set to 9. As for the "accuracy", we set the generation temperature to 0. Frozen Qwen2.5-7B is adopted as the Reader. We report the QA uncertainty and accuracy in different training strategies.
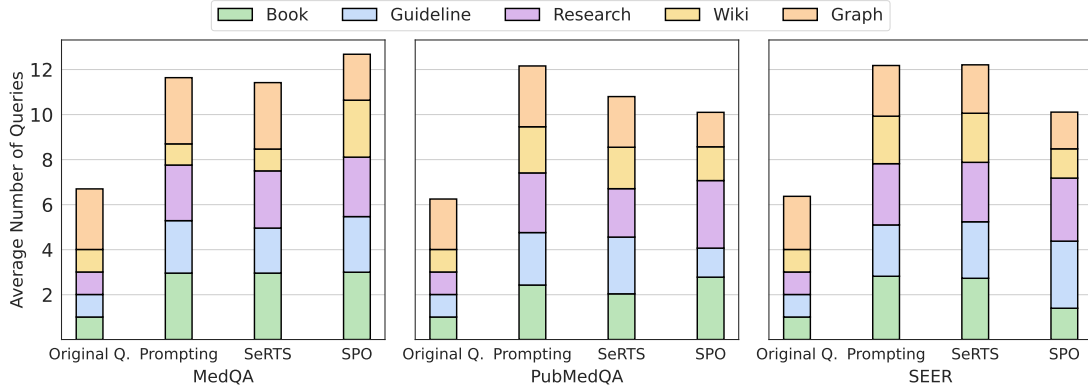
Figure 7: Average number of queries of the SPO method for different sources on test sets. "Original Q." represents "Original Question". Note that the maximum number of queries per source is 3, catering for the LLM context length limit.

Method	LLM Judger	Accuracy	Precision	Recall	F1
Trainable Planning	-	64.50	43.29	82.74	56.84
RaFe Planning	-	63.88	64.12	67.40	65.72
SPO Planning	Qwen2.5-32B	89.75	92.13	89.24	90.66
	Llama3.1-72B	90.88	90.74	92.24	91.48
	DeepSeek-R1-Distill-Llama-70B	93.50	93.06	94.81	93.93
	Qwen2.5-72B	93.00	91.67	95.19	93.40

Table 9: Human evaluation of existing methods for constructing query training data. All models are in the Instruct version and quantized using AWQ.

Training Strategy	MedQA		PubMedQA		SEER	
	Uncertainty ($\downarrow$)	Accuracy ($\uparrow$)	Uncertainty ($\downarrow$)	Accuracy ($\uparrow$)	Uncertainty ($\downarrow$)	Accuracy ($\uparrow$)
Frozen planner	0.343	64.10	0.214	53.20	0.434	53.50
+ SFT (P^{+})	0.325	74.08	0.201	59.20	0.432	58.50
+ SFT (P^{+}) + DPO ($P^{\pm}$)	**0.276**	**76.98**	**0.185**	**60.20**	**0.386**	**61.90**
+ SFT (P^{-})	0.563	61.90	0.339	46.40	0.599	52.40

Table 10: Uncertainty and accuracy of the reader paired with planners trained using different strategies.

Cases	# GPUs (# Processes)	Retrieval Time per Sample (s)	Inference Time per Sample (s)	Estimated Total Time on Training/Test Dataset (h)
Training Data Construction (24199 samples)	1	7.1	60.1	451.7
	2	4.2	30.1	230.6
	4	3.0	15.8	126.4
	8	2.8	8.3	74.6
	16	2.7	4.4	47.7
Testing Using SPO Model (8248 samples)	1	4.1	2.1	14.2
	2	2.2	1.1	7.6
	4	1.4	0.6	4.6
	8	1.2	0.3	3.4

Table 11: The detailed cost analysis for the training data construction and the test process. During training, inference involves the Qwen2.5-72B-Instruct-AWQ model generating multiple queries and judging their corresponding document sets. In testing, inference involves the trained Qwen2.5-7B-Instruct model generating multiple queries. Note that the time taken by the Reader to respond during testing is not included in this table, as it varies with the Reader's size and prompts.

As illustrated in Table 10, we can find that both phases of SPO training (SFT and DPO) reduce the uncertainty of retrieved content and improve QA accuracy. However, the uncertainty becomes higher, and QA accuracy becomes lower when the planner is trained only on negative plans. It indicates that the LLM judger has considered the ambiguous and uncertain information in the retrieval results.

A.4 Human Evaluation of Long-form Generation

Medical researchers (graduate students) in Shanghai are instructed to rate the predictions for long-form generation questions (LiveQA, MedicationQA, and ExpertQA-Biomed) using a reference solution. The annotators are informed that the annotation is conducted for academic purposes only, and they participate in the task voluntarily without special payments required. Following Wang et al. (2024c); Liao et al. (2024a), each response is scored across four aspects—Relevance, Completeness, Medical Proficiency, and Interpretability—on a grading scale from 1 to 5. The instructions are shown in Instruction E.12.

B Additional Analysis

B.1 Visualization of Planning Statistics

Figure 7 visualises the average number of queries generated by the baselines (Original Question, Prompting, and SeRTS) and our SPO method. This comparison highlights the extent to which the planners prioritise different sources. In terms of total queries, SPO generates more queries on MedQA than the baselines, reflecting the dataset's need for comprehensive knowledge as it deals with diverse topics of medical diagnosis. Conversely, for PubMedQA and SEER, which involve specialised medical knowledge, SPO generates more concise queries than the baselines. Additionally, regarding the planner's focus on sources, SPO provides a more balanced query distribution for MedQA. It allocates more focused attention to the "Research" source in PubMedQA and the "Guideline" source in SEER, aligning with their corresponding real requirements.

B.2 More Planning Statistics

We count the number of words per query and compute the semantic diversity of queries per plan. The results are averaged across all test samples. The

# Words	Book	Guideline	Research	Wiki	Graph	Average
Original	117.1	117.1	117.1	117.1	119.5	117.6
Prompting	5.7	6.5	7.0	3.9	5.8	5.8
SeRTS	6.5	6.7	7.0	4.1	5.6	6.0
SPO Planning	5.4	6.8	6.9	3.9	6.6	6.0

Table 12: Statistics of the number of words per query.

Diversity	Book	Guideline	Research	Wiki	Graph	Average
Original	-	-	-	-	0.301	0.301
Prompting	0.248	0.204	0.275	0.232	0.297	0.251
SeRTS	0.237	0.193	0.192	0.224	0.285	0.226
SPO Planning	0.220	0.173	0.218	0.207	0.287	0.221

Table 13: Diversity of queries per plan.

computation of semantic diversity follows the way in Appendix A.1.

As illustrated in Table 12, we can find that the number of words in "Original Question" is significantly higher compared to other planning methods, which hinders the acquisition of knowledge of a wide semantic range. Moreover, the SPO queries for the Wiki source have the fewest words, which is reasonable since the Wiki source contains encyclopaedic entries.

Table 13 shows that the overall semantic diversity of SPO's queries is the lowest. This is meaningful since the SPO approach can obtain more focused information based on the source's characteristics compared to baselines. The queries constructed by baselines are more dispersed due to the inability to perceive the sources.

B.3 Cost Analysis

The environmental information is listed as follows:

- **Retrieval.** No GPUs are required. (Qdrant supports CPU vector search acceleration)

- **Reranking.** Only 1 GPU is required (with a minimum of 2 GB of graphics memory needed for the efficiency of the reranking model).

- The above two modules are deployed using FastAPI, allowing MedOmniKB to handle multiple requests simultaneously.

- **Inference.** At least 1 GPUs are required. We use the 80G NVIDIA A100 to run a single LLM under the vLLM framework. (The more GPUs available, the faster SPO can construct training data and perform inference across the entire dataset)

We randomly select 1000 training samples and 1000 test samples. First, we conduct the comparison of cost with existing supervised methods, as

Method	Retrieval Time	Inference Time
Trainable Planning	7.1	64.5
RaFe Planning	7.1	2.3
SPO Planning	7.1	60.1

Table 14: Comparison of the cost of training data construction with existing supervised methods.

Category	Count
Anatomy	417
Biochemistry	1968
Ethics and Law in Medicine	254
Immunology	568
Internal Medicine	2268
Microbiology	755
Neurology	1713
Obstetrics and Gynecology	518
Pathology	1774
Pediatrics	648
Pharmacology	1194
Physiology	753
Psychiatry	1226
Public Health	1351
Radiology	1084
Surgery	1691
Total	**18182**

Table 15: Statistics of collected books.

shown in Table 14. It can be observed that our SPO method remains competitive among supervised approaches, especially given that it offers the best supervision signals, as highlighted in Section A.2.

Then we measure the time spent on retrieval and inference in various multi-process cases. We divide the total time spent by the number of samples to get the average time spent. The multi-GPU (one process on each GPU) will reduce the average time spent. Finally, we multiply the average by the total number of samples to get the estimated time on the entire training/test dataset. The cost results are shown in Table 11. The process of constructing training data is computationally intensive, but is performed only once during model training. The actual cost during real-time inference is much lower, which is suitable for real-world, real-time use cases. For the training data construction process, we recommend 1) using 4-8 GPUs to execute the process, or alternatively, 2) reducing the number of training samples. In Section 5.9, we demonstrate that the SPO method enhances planning capability even with limited training samples.

C Case Studies

Two case studies are conducted as shown in Table 16. In the first case, it is difficult to retrieve useful information from various knowledge sources with such a complex medical diagnosis as a search term. Our proposed SPO is able to propose queries that better fit the characteristics of the sources, compared to Prompting and SeRTS. Specifically, appropriate queries for the "Guideline" source are constructed by SPO, whereas queries proposed by Prompting and SeRTS are only roughly descriptive and have a high degree of duplication with other sources' queries. Moreover, for such complex problems, queries with more perspectives are generated by SPO, such as those in the "Wiki", to provide comprehensive knowledge. In the second case, where a concise biomedical question is presented, the SPO method prioritises the most relevant "Research" source. In contrast, the other two methods continue to consider the irrelevant "Guideline" source, primarily due to their limited awareness of the multi-source knowledge base.

D Additional Details

D.1 MedOmniKB Details

The category statistics of gathered books are detailed in Table 15.

For datasets that provide reference documents (MedQA, PubMedQA, BioASQ, LiveQA, MedicationQA), MedOmniKB includes either the original reference materials or similar online resources. For datasets that do not explicitly provide reference documents (MedMCQA, MMLU-Med, SEER, DDXPlus, MIMIC-IV-ED, ExpertQA-Biomed), MedOmniKB is expected to meet knowledge requirements through its extensive and diverse knowledge entries.

Importantly, this scenario simulates real-world medical scenarios, where annotated reference documents for health issues are often unavailable. The superior SPO performances on these datasets, which lack reference annotations, further demonstrate its reliability and potential for broad medical applications.

D.2 Dataset Details

The 11 datasets adopted in our work are shown in Table 3 and Table 17.

- **MedQA** (Jin et al., 2020) includes clinical medicine questions from the United States

Case Study 1	
Original Question	A 48-year-old man is brought to the emergency department for sudden onset of difficulty breathing 6 hours ago. For the past several months, he has had shortness of breath on exertion and while lying down on the bed, frequent headaches, and swelling of his feet. He does not take any medications despite being diagnosed with hypertension 10 years ago. His pulse is 90/min, respirations are 20/min, blood pressure is 150/110 mm Hg, and temperature is 37.0°C (98.6°F). Physical examination shows an overweight male in acute distress with audible wheezes. Crackles are heard bilaterally and are loudest at the lung bases. Which of the following findings on cardiac auscultation will most likely be present in this patient? A. Loud P2. B. S3 gallop. C. Absent S4. D. A loud S1.
Prompting	\<book\> Cardiac auscultation in heart failure ; Physical examination findings in heart failure ; S3 gallop in heart failure \</book\> \<guideline\> Clinical presentation of heart failure ; Auscultation findings in heart failure\</guideline\> \<research\> Auscultation findings in patients with heart failure ; S3 gallop in heart failure \</research\> \<wiki\> \</wiki\> \<graph\> heart failure , auscultation findings ; S3 gallop , clinical significance ; pulmonary edema , auscultation findings \</graph\>
SeRTS	\<book\> Cardiac auscultation in heart failure ; Physical examination findings in hypertensive heart disease ; S3 gallop in heart failure \</book\> \<guideline\> Management of hypertensive heart disease ; Clinical presentation of heart failure ; Auscultation findings in patients with hypertension and heart failure \</guideline\> \<research\> Auscultation findings in patients with hypertensive heart failure ; S3 gallop as a marker of heart failure severity \</research\> \<wiki\> \</wiki\> \<graph\> heart failure , auscultation findings ; hypertensive heart disease , cardiac auscultation \</graph\>
SPO Planning	\<book\> Cardiac auscultation in heart failure ; physical examination of heart sounds in hypertensive patients ; auscultation in patients with bilateral crackles \</book\> \<guideline\> clinical guidelines for heart failure management ; guidelines for diagnosing pulmonary edema ; cardiac auscultation in patients with heart failure \</guideline\> \<research\> Auscultation findings in patients with hypertensive heart failure ; auscultation in patients with shortness of breath \</research\> \<wiki\> cardiac auscultation ; heart failure symptoms ; pulmonary edema \</wiki\> \<graph\> heart failure , auscultation findings ; S3 gallop , clinical significance ; wheezing , auscultation findings \</graph\>
Case Study 2	
Original Question	Is the determination of specific IgE against components using ISAC 112 a reproducible technique? A. yes. B. no. C. maybe.
Prompting	\<book\> reproducibility of ISAC 112 for specific IgE determination ; validation studies of ISAC 112 ; reliability of ISAC 112 in clinical practice \</book\> \<guideline\> reproducibility of ISAC 112 in clinical guidelines ; recommendations for ISAC 112 use in allergy testing ; guidelines on the reliability of component-resolved diagnostics \</guideline\> \<research\> reproducibility of ISAC 112 in peer-reviewed studies ; validation of ISAC 112 for specific IgE ; reliability and accuracy of ISAC 112 in allergy diagnosis \</research\> \<wiki\> ISAC 112 reproducibility ; ISAC 112 validation ; ISAC 112 reliability \</wiki\> \<graph\> ISAC 112 , reproducibility ; specific IgE , component-resolved diagnostics ; allergy testing , reliability \</graph\>
SeRTS	\<book\> reproducibility of ISAC 112 ; specific IgE component testing reproducibility ; validation of ISAC 112 technique \</book\> \<guideline\> reproducibility of ISAC 112 ; specific IgE component testing ; recommendations for ISAC 112 use \</guideline\> \<research\> reproducibility studies of ISAC 112 ; validation of specific IgE component testing \</research\> \<wiki\> ISAC 112 reproducibility ; specific IgE component testing \</wiki\> \<graph\> component-resolved diagnostics , ISAC 112 ; specific IgE , component-resolved diagnostics ; allergy testing , reliability \</graph\>
SPO Planning	\<book\> reproducibility of ISAC 112 in specific IgE determination ; ISAC 112 and specific IgE testing \</book\> \<guideline\> \</guideline\> \<research\> ISAC 112 accuracy in specific IgE determination ; ISAC 112 reliability in component-resolved diagnostics ; ISAC 112 in allergy research \</research\> \<wiki\> ISAC 112 and specific IgE testing \</wiki\> \<graph\> \</graph\>

Table 16: Two case studies comparing the different source planning methods.

Dataset	Type	Medical Application
MedQA		Medical Reasoning Involving Diagnoses, Treatments, and Examinations
MedMCQA	Reasoning QA	Commonsense Reasoning Focused on Medical Knowledge
MMLU-Med		Commonsense Reasoning Focused on Medical Knowledge
PubMedQA	Research QA	Knowledge Acquisition Centered on Advanced Research
BioASQ		Knowledge Acquisition Centered on Advanced Research
SEER		Treatment Planning for Patients with Different Disease Severity
DDXPlus	Clinical QA	Diagnostic Based on Patient Dialogue Records
MIMIC-IV-ED		Clinical Outcome Prediction Using EHR Records for Decision-Making
LiveQA		Consumer Health Inquiry Support Across Various Domains
MedicationQA	Long-form Answering	Consumer Health Inquiry Support Regarding Medications
ExpertQA-Biomed		Professional-Level Biology and Healthcare Solutions

Table 17: The type and medical application of each dataset.

Medical Licensing Examination, covering topics such as diagnoses, treatments, and examinations. We adopt their 4-option English subset. In particular, 6000 training samples, 1200 development samples and 1273 test samples are selected from their official sets.

- **MedMCQA** (Pal et al., 2022) are collected from Indian medical entrance exams. Their questions cover 2400 healthcare topics and 21 medical subjects with high topical diversity. We randomly sample 6000 training samples from the official training split. The 1200 development samples and 1200 test samples are randomly selected from the official development set since their test set lacks labelled answers.

- **MMLU-Med** (Hendrycks et al., 2021) comprises six tasks related to biomedicine, which include anatomy, college biology, college medicine, clinical knowledge, human genetics and professional medicine. There are a total of 1089 test samples. MMLU-Med is included solely in our test set due to the limited number of training and development samples available in the dataset.

- **PubMedQA** (Jin et al., 2019) is a biomedical QA dataset, containing 1000 manually annotated questions based on PubMed abstracts. We remove the origin-given contexts for the RAG setting. The original 500 test samples are adopted as our test set. For the remaining samples, 417 samples are included in the training set, while 83 samples are included in the development set.

- **BioASQ** (Tsatsaronis et al., 2015) is from an annual competition for biomedical question-answering[8]. Following Xiong et al. (2024a), we select the Yes/No questions in Task B of the competition from 2014 to 2024. The questions are also based on biomedical literature. We remove the given context for the RAG setting. We randomly divide all available samples into three portions: 584 samples for the training set, 116 samples for the development set, and 782 samples for the test set.

- **SEER** (Dubey et al., 2023) acts as a treatment planning dataset involving the Surveillance, Epidemiology, and End Results (SEER) custom breast cancer databases. For each patient record, 17 attributes and recommended treatment plans are noted. We borrow the processed dataset from MedS-Bench (Wu et al., 2025), which simplifies the task into a 7-option recommendation format. We randomly divide samples into three portions: 2113 samples for the training set, 422 samples for the development set, and 1000 samples for the test set.

- **DDXPlus** (Tchango et al., 2022) is an advanced large-scale dataset designed for Automatic Symptom Detection (ASD) and Automatic Diagnosis (AD) systems. LLMs must make diagnostic decisions based on dialogues and select from 49 potential diagnoses. We also adopt the data from MedS-Bench (Wu et al., 2025). We randomly divide samples into three portions: 5000 samples for the training set, 1000 samples for the development set,

---

[8] https://www.bioasq.org/

and 1000 samples for the test set.

- **MIMIC-IV-ED** (Xie et al., 2022) is a standardised benchmark derived from the Medical Information Mart for Intensive Care IV-Emergency Department (MIMIC-IV-ED) database. Following MedS-Bench (Wu et al., 2025), three medical tasks are performed. Given a patient's EHR, LLMs are required to predict whether the patient needs hospitalisation, needs to revisit the emergency department within 72 hours, or should be classified into a critical triage queue. They can help clinicians make clinical decisions. We randomly divide samples into three portions: 2917 samples for the training set, 583 samples for the development set, and 1000 samples for the test set.

- **LiveQA** (Abacha et al., 2017) focus on answering consumer health questions received by the U.S. National Library of Medicine. Manually collected and validated reference answers from trusted sources, such as the National Institute, are provided. We adopt the 104 original test samples as the test set. The remaining samples are divided into two portions: 343 samples for the training set and 68 samples for the development set.

- **MedicationQA** (Abacha et al., 2019) is constructed by collecting anonymized consumer questions submitted to MedlinePlus[9]. Four annotators participated in the manual annotation and answering process. Finally, the answers are validated by medical experts. We randomly divide samples into three portions: 450 samples for the training set, 90 samples for the development set, and 150 samples for the test set.

- **ExpertQA-Biomed** (Malaviya et al., 2024) is a high-quality long-form QA dataset across 32 fields. The dataset comprises questions and answers proposed and verified by domain experts. Among them, 96 biology (Bio) questions and 504 medical questions (Med) are used in this work. We randomly divide these samples into three portions: 375 samples for the training set, 75 samples for the development set, and 150 samples for the test set.

[9] https://medlineplus.gov/

D.3 Baseline Details

- **No Retrieval**: The reader is prompted to think step-by-step and answer the question without external documents (Prompt E.10).

- **Original Question** (Xiong et al., 2024a): The MedRAG method, which uses the original question to retrieve a single mixed source, is adapted to our multi-source setting. Specifically, the original question is used as the query for the four unstructured sources: "Book", "Guideline", "Research" and "Wiki". For the "Graph" source, we prompt the Qwen2.5-72B-Instruct-AWQ to extract medical terms of less than 4 (Prompt E.5), which will be used as queried terms.

- **Query2Doc** (Wang et al., 2023b): We prompt the Qwen2.5-72B-Instruct-AWQ to enhance original question by producing pseudo-documents. For the four unstructured sources, we produce the documents in the corresponding style (Prompt E.6). For the structured "Graph" source, we use the same approach as in the Original Question.

- **Prompting** (Ma et al., 2023): Qwen2.5-72B-Instruct-AWQ is directly prompted to construct proper queries for each source. The number of queries for each source is limited to 0-3.

- **Reflexion** (Shinn et al., 2023): Self-reflection on a single reasoning path is conducted in this baseline. Following Hu et al. (2024), we iteratively carry out the following steps: 1. Construct appropriate queries for each source by incorporating all previous feedback (if any) (Prompt E.8). 2. Generate feedback and assign scores for the new plan (Prompt E.9). If the score reaches 5 (the maximum), the loop stops. The maximum number of iterations is set to 8 due to significant inference costs. This method is adapted to our multi-source setting by proposing and evaluating queries for all sources.

- **SeRTS** (Hu et al., 2024): Based on Reflexion, sibling nodes are added and join the reasoning process. We follow Hu et al. (2024) to perform the four operations (Selection, Expansion, Evaluation and Backpropagation) iteratively. If the score reaches 5 (the maximum),

the loop stops. The maximum number of iterations is set to 8 due to significant inference costs. This method is adapted to our multi-source setting by proposing and evaluating queries for each source.

- **Trainable Planning** (Ma et al., 2023): Different from the proposed SPO Planning, we use the performance of Qwen2.5-72B-Instruct-AWQ on the downstream tasks to measure the value of each query. The metric is the improvement of accuracy for closed-ended QA, and Rouge-L (Lin, 2004) for open-ended generation. The other SFT and DPO training processes remain the same as in SPO Planning.

- **RaFe Planning** (Mao et al., 2024): Unlike the proposed SPO Planning, we use the rerank scores between the retrieved documents and the question to assess the value of each query. The rerank score threshold is set to 2.3 by trying on the development set. The other SFT and DPO training processes remain the same as in SPO Planning.

For the baselines involving retrieval, the reader will read the retrieved documents and finally answer the question after step-by-step thinking (Prompt E.11).

D.4 Implementation Details

We employ the superior LLM, Qwen2.5-72B-Instruct (Qwen Team, 2024) to explore multiple plans and verify their correctness. The Activation-aware Weight Quantization (AWQ) (Lin et al., 2024a) accelerates inference and reduces memory requirements.

We adopt Qwen2.5-7B-Instruct as the backbone of the SPO process. Low-Rank Adaptation (LoRA) (Hu et al., 2022) is adopted in the SFT and DPO process for Parameter-Efficient Fine-Tuning. For the hyperparameters of the SFT process, the batch size is 64, the peak learning rate is 2.5e-4, and the number of epochs is 5. As for the DPO process, the batch size is 64, the peak learning rate is 5e-6, and the number of epochs is 3. We use vLLM (Kwon et al., 2023) for fast inference and set the temperature to 0 for reproducible outcomes.

E Prompt List

Prompt E.1: Description of sources in MedOmniKB

book: The API provides access to medical knowledge resources, including various educational resources and textbooks.

guideline: The API provides access to clinical guidelines from leading health organizations.

research: The API provides access to advanced biomedical research, facilitating access to specialized knowledge and resources.

wiki: The API provides access to general knowledge across a wide range of topics.

graph: The API provides a structured knowledge graph that connects medical definitions and related terms.

Prompt E.2: Query formats of sources in MedOmniKB

book: {search_query0} ; {search_query1} ; ... (Use ; to separate the queries, 0 to 3 queries)

guideline: {search_query0} ; {search_query1} ; ... (Use ; to separate the queries, 0 to 3 queries)

research: {search_query0} ; {search_query1} ; ... (Use ; to separate the queries, 0 to 3 queries)

wiki: {search_query0} ; {search_query1} ; ... (Use ; to separate the queries, 0 to 3 queries)

graph: {medical_term0} , {query_for_term0} ; {medical_term1} , {query_for_term1} ; ... (Use ; to separate the queries, 0 to 3 queries. Each query should use , to separate the {medical_term} and {query_for_term})

Prompt E.3: Planning exploration

To answer the question labeled as # Question, please construct appropriate queries to get the information you need.

1. Each source in # Source Description must have search queries constructed.
2. Please give the search queries following the format in # Query Format. Each source should have {n_queries} queries, separated by ';'. Please ensure the diversity of queries from the same source.
3. The queries for the source should accurately reflect the specific information needs from that source.

Question
{question}

Source Description
(Prompt E.1, "0 to 3 queries" is removed)

Query Format
(Prompt E.2, "0 to 3 queries" is removed)

Prompt E.4: Planning judging

You are a professional medical expert. Please judge whether the information in the # Documents supports the # Gold Answer as a response to the # Question. Please first think step-by-step and then show your judgement. Your responses will be used for research purposes only, so please have a definite answer.
You should respond to the following question using the format <answer>yes/no</answer> at the end of your response. Please keep your entire response simple and complete, up to 100 words.

Question
{question}

Gold Answer
{gold}

Documents
{documents}

Hint: Please judge whether # Documents supports the # Gold Answer in response to the # Question, rather than evaluating if the # Question's answer is the # Gold Answer.

Prompt E.5: Original Question (extract terms for querying Graph)

You are a helpful medical expert. Please return all medical terminologies in the input # Question. Each term should be split by ','.
The output format should be like:
<term>term0 , term1 , ...</term> (The number of terms should be less than 4)

Question
{question}

Prompt E.6: Query2Doc

You are a professional medical expert. Please write a passage that answers the given # Question. The passage should be considered as part of the source described in the # Source Description provided. The result should be formatted as <passage>your passage</passage>. Please keep your entire passage up to 200 words.

Question
{question}

Source Description
{source_name}: {source_desc} (For a given source, provide its source name and description, which can be found in Prompt E.1.)

Prompt E.7: Prompting, Trainable Planning, RaFe Planning, SPO Planning

To answer the question labeled as # Question, please construct appropriate queries to get the information you need.

1. Please give the search queries following the format in # Query Format. The source can have up to 3 queries, separated by ';'. Please ensure the diversity of queries from the same source. For each source, if you think no information retrieval is needed, simply output an empty tag for that source, for example: <book></book>.
2. The queries for the source should accurately reflect the specific information needs from that source.

Question
{question}

Source Description
(Prompt E.1)

Query Format
(Prompt E.2)

Prompt E.8: Reflextion and SeRTS (reflect and propose improved queries)

To answer the question labeled as # Question, please construct appropriate queries to get the information you need. Pay attention to # Feedback from previous search queries.
1. Please give the search queries following the format in # Query Format. The source can have up to 3 queries, separated by ';'. Please ensure the diversity of queries from the same source. For each source, if you think no information retrieval is needed, simply output an empty tag for that source, for example: <book></book>.
2. The queries for the source should accurately reflect the specific information needs from that source.

Question
{question}

Feedback
{feedback}

Source Description
(Prompt E.1)

Query Format
(Prompt E.2)

Please review the multiple queries and corresponding suggestions in # Feedback, and construct improved queries.

Prompt E.9: Reflextion and SeRTS (produce the feedback and score)

You are a highly intelligent agent. I am currently answering the # Question. Then, I have retrieved relevant content # Documents using corresponding ## source: xxx; query: xxx.
1. Please rate # Documents using the additive 5-point scoring system described below. Points are accumulated based on the satisfaction of each criterion:
{Five-point Rubrics}
2. Please give suggestions for constructing better queries for each source. You should also consider not using certain sources.

Question
{question}

Documents
{docs}

Source Description
{source_desc}

Please provide the suggestion using the <suggestion></suggestion> tags. You should consider enabling unused sources or abandoning used ones. Please keep the suggestion within 100 words.
Please score between 0 and 5, strictly using the aforementioned additive 5-point scoring system and the format: <score> Integer Score </score>.

<suggestion>Your suggestion here</suggestion>
<score>Your score here</score>

Prompt E.10: Reader without retrieved documents

You are a professional medical expert to answer the # Question. Please first think step-by-step and then answer the question. Your responses will be used for research purposes only, so please have a definite answer.
You should think step by step and respond in the format <answer>A/B/C/...</answer> (only one option can be chosen) at the end of your response. Please keep your entire response simple and complete, up to 100 words. (Option restrictions are removed in open-ended generation.)

Question
{question}

Prompt E.11: Reader with retrieved documents

You are a professional medical expert to answer the # Question. Please first think step-by-step using the # Retrieved Documents and then answer the question. Your responses will be used for research purposes only, so please have a definite answer.
You should think step by step and respond in the format <answer>A/B/C/...</answer> (only one option can be chosen) at the end of your response. Please keep your entire response simple and complete, up to 100 words. (Option restrictions are removed in open-ended generation.)

Retrieved Documents
{documents}

Question
{question}

Instruction E.12: Instruction for human evaluation for long-form answering

Relevance:
1: Completely unrelated to the question
2: Some relation to the question, but mostly off-topic
3: Relevant, but lacking focus or key details
4: Highly relevant, addressing the main aspects of the question
5: Directly relevant and precisely targeted to the question

Completeness:
1: Extremely incomplete
2: Almost incomplete with limited information
3: Moderate completeness with some information
4: Mostly complete with most of the information displayed
5: Fully complete with all information presented

Proficiency in medicine:
1: Using plain language with no medical terminology.
2: Equipped with some medical knowledge but lacking in-depth details
3: Conveying moderately complex medical information with clarity
4: Showing solid grasp of medical terminology but having some minor mistakes in detail
5: Fully correct in all presented medical knowledge

Interpretability (focused on retrieved documents):
1: Minimal relevant information; does not assist in answering the question
2: Limited relevant content; offers little help in answering the question
3: Some relevant information; may need further clarification to answer the question
4: Most relevant information presented; generally aids in answering the question but lacks some details
5: Comprehensive and clear; fully supports answering the question

Question
{question}

Reference Answer
{reference}

Predicted Answer
{prediction}