

R2MED: A Benchmark for Reasoning-Driven Medical Retrieval

Anonymous ACL submission

Abstract

Current medical retrieval benchmarks primarily emphasize lexical or shallow semantic similarity, overlooking the reasoning-intensive demands that are central to clinical decision-making. In practice, physicians often retrieve authoritative medical evidence to support diagnostic hypotheses. Such evidence typically aligns with an inferred diagnosis rather than the surface form of a patient’s symptoms, leading to low lexical or semantic overlap between queries and relevant documents. To address this gap, we introduce **R2MED**, the first benchmark explicitly designed for reasoning-driven medical retrieval. It comprises 876 queries spanning three tasks: Q&A reference retrieval, clinical evidence retrieval, and clinical case retrieval. These tasks are drawn from five representative medical scenarios and twelve body systems, capturing the complexity and diversity of real-world medical information needs. We evaluate 15 widely-used retrieval systems on R2MED and find that even the best model achieves only 31.4 nDCG@10, demonstrating the benchmark’s difficulty. Although reasoning-based enhancements improve performance, a large gap remains. These findings underscore a substantial gap between current retrieval techniques and the reasoning demands of real clinical tasks. We release R2MED as a challenging benchmark to foster the development of next-generation medical retrieval systems with enhanced reasoning capabilities.

1 Introduction

Medical information retrieval (MIR) plays an important role in clinical decision-making by helping clinicians locate relevant evidence from sources such as biomedical literature, knowledge bases, and clinical records (Luo et al., 2008; Goeriot et al., 2016; Frisoni et al., 2022; Zhang et al., 2026). In real-world settings, however, medical retrieval often requires more than lexical or semantic matching (Lee et al., 2019; Jin et al., 2023). Relevance

may depend on implicit clinical reasoning, such as inferring latent symptom–disease associations or identifying evidence that supports a diagnostic conclusion not explicitly stated in the query. Existing MIR benchmarks (Boteva et al., 2016; Voorhees et al., 2021; Li et al., 2024) largely do not capture this challenge. For example, benchmarks such as NFCorpus (Boteva et al., 2016) align queries and documents through explicit links, so lexical overlap becomes a strong signal of relevance (Figure 1 (1)). As a result, these benchmarks mainly reward shallow matching and fail to reflect the reasoning-intensive retrieval needs of clinical practice.

This limitation has become more important in modern medical question answering (Jin et al., 2021; Zuo et al., 2025; Qiu et al., 2025). Retrieval-augmented generation (RAG) depends on retrieving evidence that genuinely supports answer generation, while large reasoning models (LRMs) place increasing emphasis on multi-step clinical reasoning (Xiong et al., 2024a; Wu et al., 2024; Xiong et al., 2024b; Jaech et al., 2024; Chen et al., 2024c; Zhang et al., 2025). In both settings, useful evidence is often related to the query through implicit reasoning rather than direct surface overlap. This creates a clear need for benchmarks that explicitly evaluate reasoning-intensive medical retrieval.

In this work, we introduce R2MED, the first benchmark explicitly designed for reasoning-intensive medical retrieval. Unlike prior datasets that emphasize direct lexical overlap or shallow semantic similarity, R2MED evaluates whether a system can retrieve documents that follow the implicit reasoning path needed to answer a query (Figure 1 (1)). R2MED contains 876 queries spanning three reasoning-centric retrieval tasks and eight datasets, covering diverse clinical scenarios. Across these tasks, relevance is defined by whether a document supports the inferred diagnosis, conclusion, or clinical reasoning path behind the query, rather than by explicit query–document matching alone.

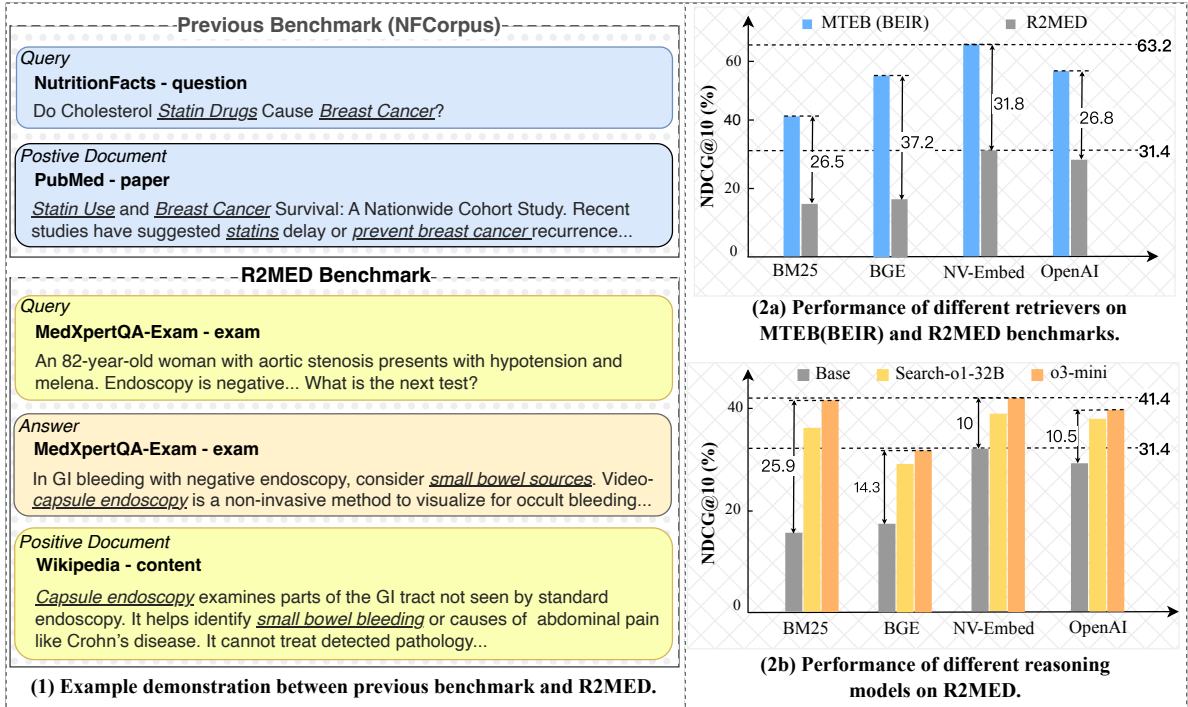


Figure 1: Overview of R2MED. Subfigure (1) contrasts R2MED with NFCorpus, showing a shift from semantic matching to reasoning-driven retrieval. Subfigures 2(a) and 2(b) report retrieval and reasoning performance on R2MED, revealing the limitations of existing retrievers under reasoning-intensive settings.

085 We benchmark 15 classical retrieval systems
086 on R2MED and observe a substantial drop in
087 performance relative to standard retrieval bench-
088 marks. For example, although NV-Embed-v2 (Lee
089 et al., 2024) reaches 63.2 nDCG@10 on the
090 MTEB (Muennighoff et al., 2022) retrieval subset
091 of BEIR (Thakur et al., 2021), it achieves only 31.4
092 nDCG@10 on R2MED (Figure 1 (2a)). We also
093 study reasoning-augmented retrieval with LRMs
094 and find that reasoning can improve performance
095 on complex medical queries, but only to a lim-
096 ited degree. Even the strongest configuration, NV-
097 Embed-v2 augmented with o3-mini reasoning guid-
098 ance, reaches only 41.4 nDCG@10 (Figure 1 (2b)).
099 These results indicate that reasoning-intensive med-
100 ical retrieval remains an open challenge.

101 **Contributions:** (1) we introduce R2MED, the
102 first benchmark for reasoning-intensive medical re-
103 trieval across three tasks and eight datasets; (2) we
104 provide a comprehensive evaluation of 15 retrieval
105 systems, showing that strong retrievers still per-
106 form poorly on R2MED, with the best nDCG@10
107 only reaching 31.4; and (3) we analyze reasoning-
108 augmented retrieval with large reasoning models,
109 showing that although reasoning improves retrieval,
110 current methods remain insufficient for the de-
111 mands of complex medical retrieval scenarios.

2 Related Work

112 **Medical Retrieval Benchmarks.** To support the
113 advancement of medical information retrieval, a
114 range of domain-specific benchmarks have been
115 developed. Most existing benchmarks like NFCor-
116 pus (Boteva et al., 2016), SciFact (Wadden et al.,
117 2020), TREC-COVID (Voorhees et al., 2021), and
118 CMIRB (Li et al., 2024) primarily focus on key-
119 words or shallow semantic matching between the
120 query and relevant documents. For instance, NF-
121 Corpus aligns layperson health questions with sci-
122 entific articles from NutritionFacts.org, using cu-
123 rated links to PubMed literature to establish rele-
124 vance. Closest to our work, BRIGHT (Su et al.,
125 2024) begins to explore reasoning-based retrieval
126 by constructing a large-scale dataset of user queries
127 paired with relevant web documents, primarily
128 sourced from community QA forums. However,
129 we take a different perspective by constructing
130 retrieval tasks grounded in authentic clinical sce-
131 narios that inherently require multi-step medical
132 reasoning. R2MED is a benchmark dedicated to
133 reasoning-centric medical retrieval, in which rel-
134 evant documents are often connected to queries
135 through complex reasoning.
136

137 **Medical QA Benchmarks.** Early medical QA

138 benchmarks such as MedQA (Jin et al., 2021),
139 MedMCQA (Pal et al., 2022), and MMLU (Med-
140 ical) (Hendrycks et al., 2020) are primarily de-
141 rived from medical licensing and entrance exami-
142 nations. These benchmarks focus on basic medical
143 knowledge understanding in standardized, multiple-
144 choice formats. Recently, some work has shifted
145 focus toward clinical reasoning and complex medi-
146 cal QA. MedXpertQA (Zuo et al., 2025) presents
147 complex, specialty-specific multiple-choice ques-
148 tions grounded in real clinical settings. MedR-
149 Bench (Qiu et al., 2025) constructs open-ended
150 diagnostic and therapeutic reasoning tasks derived
151 from curated patient case reports. These bench-
152 marks reflect a growing emphasis on robust, multi-
153 step clinical reasoning in medical QA. We curate
154 retrieval-focused queries from a subset of these
155 complex QA datasets and enhance them with addi-
156 tional annotations to construct R2MED.

Dense Retrieval. Modern information retrieval
157 has evolved significantly with the rise of dense
158 retrieval models, which encode queries and docu-
159 ments into continuous vector spaces. Representa-
160 tive models such as Contriever (Izacard et al.,
161 2021), BGE (Xiao et al., 2024), BMRetriever (Xu
162 et al., 2024), and GritLM (Muennighoff et al.,
163 2024) are typically pre-trained on large-scale cor-
164 pora and further fine-tuned using supervised or
165 synthetic data. Generation-augmented retrieval
166 methods such as HyDE (Gao et al., 2022) and
167 Query2doc (Wang et al., 2023b) further reduce the
168 semantic gap by synthesizing hypothetical docu-
169 ments. More recent work incorporates explicit rea-
170 soning into retrieval, such as inference-time logical
171 reasoning (Faltings et al., 2025) and logic-aware
172 multi-hop retrieval frameworks like HopRAG (Liu
173 et al., 2025a). In parallel, large reasoning mod-
174 els such as o1 (Jaech et al., 2024) and DeepSeek-
175 R1 (Guo et al., 2025), as well as agentic search sys-
176 tems like Search-o1 (Li et al., 2025b) and Search-
177 r1 (Jin et al., 2025), have shown promise on com-
178 plex reasoning-intensive retrieval tasks. In this
179 work, we evaluate these emerging paradigms un-
180 der a unified reasoning-centric medical retrieval
181 setting, revealing their strengths and limitations.

183 3 R2MED: A New Reasoning-Driven 184 Retrieval Benchmark

185 3.1 Preliminary

186 Reasoning-driven medical information retrieval
187 poses unique challenges that go beyond surface-

188 level lexical or semantic matching. Formally,
189 given a query q and a document corpus $\mathcal{D} =$
190 $\{d_1, \dots, d_n\}$, the task is to identify a subset of rel-
191 evant documents $\mathcal{D}_q^+ = \{D_{q,1}^+, \dots, D_{q,m}^+\} \subset \mathcal{D}$,
192 where $m \ll n$. All remaining documents are
193 treated as negative examples, denoted by $\mathcal{D}_q^- =$
194 $\mathcal{D} \setminus \mathcal{D}_q^+$. Unlike conventional retrieval tasks, rel-
195 evance in this context is mediated by a latent rea-
196 soning answer \mathcal{A} that logically links the query
197 to its corresponding positive documents. Import-
198 antly, this reasoning answer is often absent from
199 the query’s surface form, requiring models to infer
200 it implicitly via reasoning.

201 3.2 Task Curation

202 R2MED is a benchmarking dataset designed to
203 evaluate retrieval systems in reasoning-intensive
204 medical scenarios. It comprises three tasks: Q&A
205 reference retrieval, clinical evidence retrieval, and
206 clinical case retrieval, each targeting a distinct type
207 of clinical information need (see Table 1).

208 The Q&A reference retrieval task aims to re-
209 trieve high-quality external resources that provide
210 essential evidence for answering medical questions.
211 Each query is a natural language question sourced
212 from a community post on the StackExchange plat-
213 form. Relevant documents refer to webpages cited
214 within the corresponding answer, having under-
215 gone expert validation to ensure they convey criti-
216 cal knowledge essential for answering the question.
217 Therefore, the answer serves as an implicit reason-
218 ing anchor that links the question to its relevant
219 documents.

220 The clinical evidence retrieval task focuses on re-
221 trieving medical evidence that supports diagnostic
222 or treatment planning within the clinical decision-
223 making scenario. Each query is a complex clinical
224 question drawn from established medical QA
225 datasets. Relevant documents are curated from au-
226 thoritative medical encyclopedias and verified to
227 provide sufficient evidence for the clinical decision
228 implied by the query. The original answer in the
229 QA dataset thus serves as an implicit reasoning step
230 that bridges the query and its relevant documents.

231 The clinical case retrieval task centers on re-
232 trieving cases with the same diagnosis to assist in
233 analyzing a given patient scenario. Each query is a
234 structured clinical description, including chief com-
235 plaint, history, and physical findings, sourced from
236 case reports or electronic health records. Relevant
237 documents are clinical cases sharing the same diag-
238 nosis and verified to provide support for the query.

Dataset	#Q	#D	Avg. Pos	Q-Len	D-Len	Q-Source	D-Source	Example
Q&A Reference Retrieval Task								
Biology	103	57,359	3.6	115.2	83.6	StackExchange post (Stack Exchange Network, 2026)	Web pages: article, blog, wikipedia ...	Tab. 15
Bioinformatics	77	47,473	2.9	273.8	150.5			Tab. 16
Medical Sciences	88	34,810	2.8	107.1	122.7			Tab. 17
Clinical Evidence Retrieval Task								
MedXpertQA-Exam	97	61,379	3.0	233.2	154.9	Exam question (Zuo et al., 2025)	Wikipedia (Xiong et al., 2024a)	Tab. 18
MedQA-Diag	118	56,250	4.4	167.8	179.7	Exam question (Jin et al., 2021)	Textbooks (Jin et al., 2021)	Tab. 19
PMC-Treatment	150	28,954	2.1	449.3	149.3	Clinical question (Qiu et al., 2025)	PubMed articles (Qiu et al., 2025)	Tab. 20
Clinical Case Retrieval Task								
PMC-Clinical	114	60,406	2.2	182.8	480.4	Clinical Case (Zhao et al., 2023)	PubMed cases (Zhao et al., 2023)	Tab. 21
IYi-Clinical	129	10,449	3.5	602.3	1,273.0	Clinical Case (IYi, 2026)	IYi cases	Tab. 22

Table 1: Statistics of R2MED. #Q and #D denote the number of queries and documents, respectively. Avg. Pos refers to the average positive documents per query. Q-Len and D-Len are the average lengths of queries and documents. We measure the average length by the GPT-2 tokenizer (Radford et al., 2019).

Here, the diagnosis serves as a latent reasoning bridge linking the query to its relevant documents.

3.3 Benchmark Construction

Data Collection. As illustrated in Figure 2, our dataset construction begins with a systematic and task-specific collection process grounded in high-quality medical corpora. R2MED comprises eight datasets drawn from diverse sources, reflecting variations in data modalities (Table 1). At this stage, we curate a unified quadruple $(Q, A, D_{init}^+, D_{init}^-)$ for each dataset, representing the query, gold answer, initial positive documents, and initial negative documents, respectively. For the Q&A reference retrieval task, we curate query-answer pairs (Q, A) from three StackExchange communities, namely Biology, Bioinformatics, and Medical Sciences, by selecting posts with accepted or highly upvoted answers. The webpages linked within these answers form D_{init}^+ , while negative documents D_{init}^- are sampled from Wikipedia. Among these, the Biology dataset is adopted directly from the BRIGHT benchmark (Su et al., 2024). For the clinical evidence retrieval task, we reformat three medical QA datasets (MedXpertQA (Zuo et al., 2025), MedQA (Jin et al., 2021), MedR-Bench (Qiu et al., 2025)) into three specialized datasets, each corresponding to a different stage of clinical decision-making: examination recommendation, diagnosis, and treatment planning. Candidate documents D are drawn from three high-quality sources: Wikipedia (Xiong et al., 2024a), medical textbooks (Jin et al., 2021), and PubMed articles (Qiu et al., 2025). For the clinical case retrieval task, we collect full patient records from PMC-Patients (Zhao et al., 2023) and the IYi-bingli website. We extract the structured clinical

presentation as Q , and the confirmed diagnosis as A from each record by GPT-4o. Clinical cases with the same diagnosis form D_{init}^+ , while other cases form D_{init}^- . We also apply a series of filtering and restructuring steps to ensure that the resulting queries align with the intended retrieval tasks. Please refer to Appendix A.1 for further details.

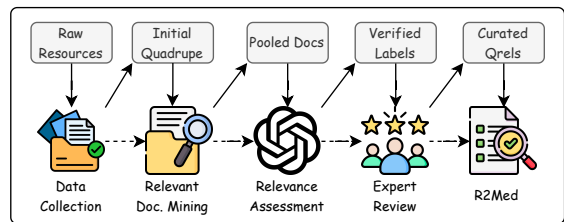


Figure 2: R2MED benchmark construction pipeline.

Relevant Document Mining. While each dataset in R2MED is initially constructed with a quadruple $(Q, A, D_{init}^+, D_{init}^-)$, the negative set D_{init}^- may contain false negatives that are relevant but unverified (Chen et al., 2024b; Moreira et al., 2024). To enrich the positive document pool and mitigate noise in the negatives, we adopt a retrieval-based mining strategy. For each pair (q, a) , we use OpenAI o3 model to generate a reasoning path s , forming a multi-view retrieval set $\mathcal{S}_q = \{q, a, s\}$. To ensure retrieval diversity, we employ a retrieval committee $\mathcal{C} = \{r_1, r_2, \dots, r_n\}$, where each r_i denotes a distinct retriever. For each element in \mathcal{S}_q , each $r_i \in \mathcal{C}$ independently retrieves top- k documents from D_{init}^- . We aggregate the retrieved results from all committee members and rank candidate documents based on their frequency of appearance. The top- k most frequently retrieved documents are selected as the mined relevant set $D_{q,ret}$, which is merged with the initial positives to form the

enhanced positive pool $\mathcal{D}_{q,\text{pool}}^+$. Simultaneously, these documents are removed from $\mathcal{D}_{q,\text{init}}^-$ to update the negative pool $\mathcal{D}_{q,\text{pool}}^-$. An intermediate quadruple $(\mathcal{Q}, \mathcal{A}, \mathcal{D}_{\text{pool}}^+, \mathcal{D}_{\text{pool}}^-)$ is yielded in this process. See Appendix A.2 for more details.

Relevance Assessment. To ensure data quality, we perform a fine-grained relevance assessment on the pooled document sets using GPT-4o. For each candidate document $d \in \mathcal{D}_{q,\text{pool}}^+$, we evaluate its relevance using the triple (q, a, s) . The assessment follows a two-dimensional scoring rubric on a 0–10 scale, assessing i) the document’s relevance to the answer, ii) its support for the reasoning process. Documents scoring at least 8 in both dimensions are retained as verified positives $\mathcal{D}_{q,\text{ver}}^+$. Those receiving ambiguous scores (5–7) are discarded to avoid introducing noise into the evaluation. Documents scoring 4 or below are treated as verified negatives and added to the set $\mathcal{D}_{q,\text{ver}}^-$. This procedure yields a refined and rigorously validated dataset $(\mathcal{Q}, \mathcal{A}, \mathcal{D}_{\text{ver}}^+, \mathcal{D}_{\text{ver}}^-)$. Full details of the scoring protocol are provided in Appendix A.3.

Expert Review. To ensure clinical validity and factual reliability, especially in light of the involvement of language models in data generation and assessment, we conduct a final expert review stage. In this stage, a medically trained reviewer examines all data samples to identify potential quality issues. A board-certified medical expert then re-examines the flagged cases to make the final judgment. Each data point is reviewed across three criteria: (1) whether the reformulated query (if applicable) is clinically coherent and complete; (2) whether the reasoning path reflects plausible and accurate medical inference; and (3) whether the positive documents provide essential support for both the query and answer. Data that fail to meet these criteria are excluded from the final release. An additional expert agreement study shows high consistency between the final qrels and majority expert judgments (88.0% agreement, Fleiss’ $\kappa = 0.828$); details are provided in Appendix A.5.

3.4 Diversity Analysis

We assess the diversity of R2MED from both clinical and distributional perspectives. Each query is categorized by its medical scenario and involved body system. As shown in Figure 15, R2MED covers 5 major clinical scenarios and 12 body systems, capturing a wide range of real-world medical contexts. In addition, we compute weighted Jaccard

similarity across datasets and observe consistently low overlap, indicating that R2MED presents a challenging testbed requiring strong generalization across diverse and out-of-distribution domains. See Appendix C for more details.

4 Experiments

4.1 Experimental Setup

We evaluate 15 representative retrieval models, including both sparse retrieval (BM25 (Robertson et al., 2009)) and dense retrieval models (top performers on the MTEB leaderboard (Contributors, 2026)). Dense retrieval models are divided into two categories: base-size models (< 1B) such as Contriever (Izacard et al., 2021), MedCPT (Jin et al., 2023), InstructOR-L (Su et al., 2022), BGE-Large (Xiao et al., 2024), and BMRetriever-410M (Xu et al., 2024), and large-size models (> 1B) including InstructOR-XL (Su et al., 2022), BMRetriever-2B/7B (Xu et al., 2024), E5-Mistral (Wang et al., 2023a), GritLM-7B (Muennighoff et al., 2024), SFR-Embedding-Mistral (Meng et al., 2024), NV-Embed (Lee et al., 2024). We additionally evaluate two proprietary embedding models from OpenAI (OpenAI) and Voyage (Voyage AI). Among these, MedCPT and the BMRetriever family are domain-specific retrievers pretrained on large-scale biomedical corpora. Detailed model descriptions are provided in Appendix E.2. Following prior work (Nguyen et al., 2016; Thakur et al., 2021; Su et al., 2024), we use nDCG@10 as the primary evaluation metric.

4.2 Main Results

Existing retrieval systems perform poorly on R2MED. As shown in Table 2, retrieval models across a wide range of sizes and architectures achieve uniformly low performance on R2MED, with the best-performing model (NV-Embed-v2) reaching only 31.43 nDCG@10. These retrievers are primarily trained on conventional semantic relevance datasets, rendering them ineffective for reasoning-intensive retrieval. Notably, BM25 performs on par with base-size dense retrievers, while large-size models (> 1B) consistently outperform smaller ones. Interestingly, medical retrievers such as BMRetriever-7B show no clear advantage over general-purpose retrievers like GritLM-7B or NV-Embed-v2, despite pretraining on large biomedical corpora. This may stem from differences in backbone architectures as well as the lim-

Task	Size	Q&A Reference			Clinical Evidence			Clinical Case		Avg.
		Biology	Bioin.	MedS.	MedE.	MedD.	PMCT.	PMCC.	IYiC.	
Sparse Retrieval										
BM25 (Robertson et al., 2009)	-	19.19	21.55	19.68	0.66	2.55	23.69	21.66	12.02	15.13
Base Size (< 1B)										
Contriever (Izcard et al., 2021)	110M	9.15	18.02	25.22	1.71	2.52	11.47	13.40	12.57	11.76
MedCPT [†] (Jin et al., 2023)	220M	2.15	17.57	14.74	1.68	2.02	11.33	14.62	8.03	9.02
InstructOR-L (Su et al., 2022)	335M	15.82	29.71	36.88	3.84	4.81	15.84	9.02	13.77	16.21
BGE-Large (Xiao et al., 2024)	335M	12.71	27.04	27.76	4.10	8.33	26.45	15.06	14.72	17.02
BMRetriever [†] (Xu et al., 2024)	410M	12.37	29.92	31.26	4.46	6.28	25.31	17.46	17.73	18.10
Large Size (> 1B)										
InstructOR-XL (Su et al., 2022)	1.5B	21.56	32.91	36.79	4.63	4.29	14.18	14.49	16.17	18.13
BMRetriever-2B [†] (Xu et al., 2024)	2B	19.50	33.30	39.45	9.97	9.31	38.01	25.65	22.30	24.69
E5-mistral (Wang et al., 2023a)	7B	18.81	42.86	41.77	6.70	11.54	23.58	31.17	22.93	24.92
BMRetriever-7B [†] (Xu et al., 2024)	7B	23.62	44.01	44.91	11.55	16.95	<u>46.88</u>	29.14	<u>24.36</u>	30.18
SFR-Embedding (Meng et al., 2024)	7B	19.56	<u>45.91</u>	<u>46.01</u>	<u>11.98</u>	<u>17.49</u>	44.19	36.36	23.71	30.65
GritLM-7B (Muennighoff et al., 2024)	7B	24.99	43.98	45.94	12.32	19.86	39.88	<u>37.08</u>	24.94	<u>31.12</u>
NV-Embed-v2 (Lee et al., 2024)	7B	27.15	50.10	47.81	10.90	16.72	44.05	39.91	14.81	31.43
Voyage-3 (Voyage AI)	-	<u>25.42</u>	38.98	41.63	8.74	9.36	45.28	28.68	20.64	27.34
OpenAI-3-large (OpenAI)	-	23.82	40.51	44.05	11.78	15.01	47.43	28.87	17.12	28.57

Table 2: The performance of retrieval models on R2MED. We report nDCG@10 for eight datasets: Biology, Bioinformatics (Bioin.), Medical Sciences (MedS.), MedXpertQA-Exam (MedE.), MedQA-Diag (MedD.), PMC-Treatment (PMCT.), PMC-Clinical (PMCC.), IYi-Clinical (IYiC.). † denotes medical retrievers. **Bold** and underline indicate the best and second-best results on each dataset.

itations of medical training corpora, which often lack reasoning-driven retrieval data. These results underscore the limitations of current retrieval systems in complex medical contexts and motivate the development of models better aligned with the demands of reasoning-driven retrieval.

Reranking helps but does not solve R2MED.

Reranking has been widely adopted to improve retrieval performance (Nogueira et al., 2019, 2020; Liu et al., 2025b). We evaluate three representative standard rerankers, namely MonoBERT (Nogueira et al., 2019), BGE-Reranker-v2-m3 (Chen et al., 2024a), and RankLlama-7B (Ma et al., 2024), on the top-10 and top-100 documents retrieved by three retrievers, and include RankR1-14B as a reasoning-oriented top-10 reranker. Standard rerankers help weaker candidate sets, e.g., RankLLaMA top-100 improves BM25/BGE-Large to 21.74/22.54 nDCG@10, but are unstable on NV-Embed-v2, whose best standard-reranker result is 29.93 versus 31.43 without reranking. In contrast, our RankR1-14B run improves all three retrievers, reaching 20.45/23.37/36.14 on BM25/BGE-Large/NV-Embed-v2. This suggests that reasoning-oriented rerankers better capture the latent reasoning. A recent external study provides a broader

reasoning-reranker comparison on R2MED: while these models improve substantially over standard rerankers, the best reported result is still 42.85 average nDCG@10, leaving large room for stronger reasoning-based reranking. We include these results as complementary evidence; see Appendix G.

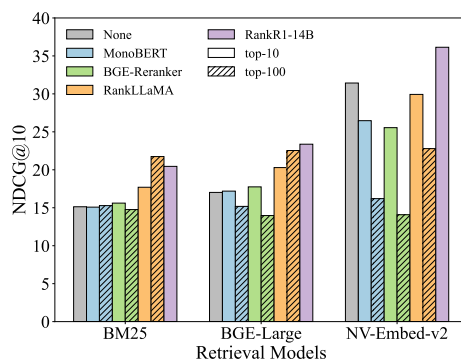


Figure 3: Average reranking performance on R2MED, including RankR1-14B. Standard-reranker details are in Table 25.

GAR methods demonstrate effectiveness on R2MED. Recently, generation-augmented retrieval (GAR) methods, enhancing queries by leveraging LLMs to generate rewritten queries or hypo-

Model	Biology	Bioin.	MedS.	MedE.	MedD.	PMCT.	PMCC.	IYiC.	Avg.
NV-Embed-v2	27.15	50.1	47.81	10.90	16.72	44.05	39.91	14.81	31.43
Large Language Model									
Qwen2.5-7B-Ins.	30.12	49.95	49.39	13.37	19.49	42.99	38.36	17.86	32.69
Qwen2.5-32B-Ins.	31.34	52.35	49.76	16.40	22.77	45.31	43.40	21.35	35.34
Llama3.1-70B-Ins.	31.21	52.27	51.19	17.48	27.53	46.96	46.90	21.05	36.82
GPT-4o	33.61	54.15	50.83	23.08	36.09	47.35	48.51	21.30	39.37
Large Reasoning Model									
R1-Distill-Qwen-32B	33.1	51.82	49.39	18.78	27.38	45.94	42.16	20.05	36.08
QwQ-32B	32.26	52.43	49.91	21.08	31.29	46.14	41.06	20.38	36.82
R1-Distill-Llama-70B	32.83	53.31	50.32	22.98	33.78	47.04	46.53	21.35	38.52
HuatuoGPT-o1-70B	31.25	52.81	49.55	25.25	38.33	48.93	48.57	21.77	39.56
o3-mini	34.01	55.9	51.28	28.99	40.30	48.97	50.86	20.47	41.35
Search-Enhanced Large Reasoning Model									
Search-R1-3B	25.76	47.53	47.57	11.98	18.88	45.66	38.57	17.95	31.74
Search-R1-7B	30.84	50.66	49.07	15.05	20.46	47.36	45.49	21.95	35.11
Search-o1 _{QwQ-32B}	31.82	53.33	51.32	21.68	32.80	45.93	47.37	21.52	38.22
Search-o1 _{Qwen3-32B}	33.46	51.04	50.20	22.91	32.02	46.88	46.18	21.32	38.00

Table 3: LRM performance on R2MED. Rows of the same color indicate shared backbones (e.g., R1-Distill-Qwen-32B uses Qwen2.5-32B-Ins). See Tables 29–31 for results with other retrievers.

thetical documents before retrieval, have emerged as a promising approach for adapting retrieval models to out-of-domain scenarios (Mao et al., 2020, 2024; Li et al., 2025a). We evaluate three representative GAR methods: HyDE (Gao et al., 2022), Query2Doc (Wang et al., 2023b), and LameR (Shen et al., 2023), each instantiated with three backbones of increasing capacity: Qwen2.5-7B-Instruct (Team, 2024), Qwen2.5-72B-Instruct (Team, 2024), and GPT-4o. As shown in Table 4, larger generators consistently yield better retrieval performance, with GPT-4o achieving the highest scores across all three methods. Notably, Query2Doc with GPT-4o delivers the highest nDCG@10 of 41.66, significantly outperforming the best vanilla retriever. BM25 benefits most from GAR approaches, possibly due to its flexibility in handling out-of-distribution queries generated by LLMs. Overall, these results reinforce a central insight of R2MED that an intermediate answer serves as a crucial semantic bridge, effectively narrowing the gap between queries and relevant documents.

5 Analysis

5.1 LRMs Bring Marginal Gains on R2MED

Recent advancements in large reasoning models (LRMs), such as OpenAI’s o1 (Jaech et al., 2024) and DeepSeek-R1 (Guo et al., 2025), have demon-

Methods	BM25	BGE.	NV.
None	15.13	17.02	31.43
HyDE _{Qwen-7B}	26.38	24.00	32.69
HyDE _{Qwen-72B}	32.07	26.89	37.00
HyDE _{GPT-4o}	37.70	28.63	39.37
Query2Doc _{Qwen-7B}	25.73	23.45	32.52
Query2Doc _{Qwen-72B}	32.43	29.19	37.79
Query2Doc _{GPT-4o}	35.85	30.72	41.66
LameR _{Qwen-7B}	25.32	25.44	34.52
LameR _{Qwen-72B}	31.90	29.87	38.56
LameR _{GPT-4o}	32.07	30.74	38.10

Table 4: Average nDCG@10 score of GAR methods. **Bold** indicates the best results on each retriever. Detailed scores can be found in Table 26 to 28.

strated strong performance on complex medical reasoning tasks (Xie et al., 2024; Jiang et al., 2025). To assess their utility for reasoning-driven retrieval, we evaluate two paradigms: LRMs and search-enhanced LRMs. The LRM group includes DeepSeek-R1-Distill-Qwen-32B (Guo et al., 2025), QwQ-32B (Team, 2025b), DeepSeek-R1-Distill-Llama-70B (Guo et al., 2025), HuatuoGPT-o1-70B (Chen et al., 2024c), and o3-mini (openai, 2025). Search-enhanced LRMs incorporate agentic search workflows that enable dynamic retrieval of external knowledge during inference, particularly when the model encounters uncertainty. We evaluate Search-R1 (Jin et al., 2025), which is based

on Qwen2.5-3b-it-em-ppo and Qwen2.5-7b-it-em-ppo, and Search-o1 (Li et al., 2025b), implemented with QwQ-32B and Qwen3-32B (Team, 2025a) as backbones. We use MedCorp (Xiong et al., 2024a) only during search-enhanced query rewriting; final retrieval remains on the same R2MED corpus, qrels, metrics, and retriever backend. All models are evaluated under the HyDE setup, while only the final answer (excluding the reasoning trace) is extracted as the rewritten query. More details are provided in Appendix E.3.

Experimental results in Table 3 show that LRMs outperform their base LLM counterparts across backbones. For example, DeepSeek-R1-Distill-Llama-70B achieves an nDCG@10 of 38.52, surpassing Llama3.1-70B-Instruct’s 36.82. This trend holds across other model pairs, indicating that enhanced reasoning capabilities contribute modestly to improved retrieval performance on R2MED. Meanwhile, fine-tuned models on medical (e.g., HuatuoGPT-o1) also show slight gains. Notably, search-enhanced LRMs bring gains by incorporating external knowledge during inference, for instance, Search-o1_{QwQ-32B} improves upon its base model QwQ-32B, raising nDCG@10 from 36.82 to 38.22. Search-o1 outperforms Search-R1 across multiple metrics, likely due to the incorporation of a reason-in-documents module that better utilizes retrieved content. Significance analysis for QwQ-32B is provided in Appendix E.4.

Despite these improvements, the overall gains remain modest, suggesting current LRMs have yet to fully realize their potential in reasoning-based retrieval. Additionally, these methods raise substantial efficiency concerns. LRMs generate long reasoning traces that increase token usage and latency, while search-enhanced models add computational overhead through multiple retrievals during generation. As such, it is crucial to assess LRMs through a balanced lens of both effectiveness and efficiency. Designing methods that jointly optimize for both remains an open and pressing challenge.

5.2 Accurate Reasoning Improves Retrieval

To gain deeper insights into how LRMs contribute to retrieval improvements on R2MED, we investigate the relationship between the accuracy of generated intermediate answers and the final retrieval performance. We focus on five datasets (excluding the Q&A reference retrieval task), as these contain well-defined medical entities or concise phrases as golden answers, thereby allowing for more reliable

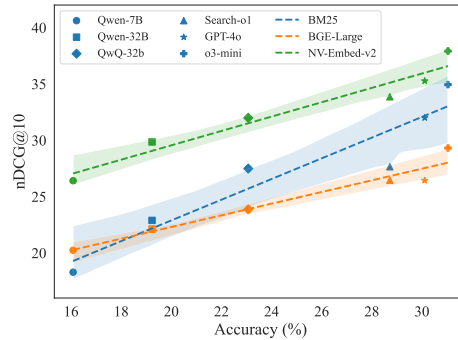


Figure 4: Correlation between reasoning answer accuracy and retrieval performance.

evaluation. We evaluate six representative models: Qwen2.5-7B-Instruct, Qwen2.5-32B-Instruct, QwQ-32B, Search-o1_{QwQ-32B}, GPT-4o, and o3-mini. For each model, we extract the predicted answer entity from its generated reasoning trace and assess its correctness using GPT-4o as a judge.

As shown in Figure 4, answer accuracy is strongly correlated with retrieval performance. Models that generate more accurate intermediate answers retrieve more relevant documents. Moreover, LRMs outperform size-matched LLMs in both answer accuracy and retrieval effectiveness, highlighting the benefits of long-chain reasoning. These results reinforce the reasoning-centric nature of R2MED, where higher answer accuracy directly contributes to better retrieval performance.

6 Conclusion and Future Work

We introduce R2MED, the first benchmark specifically designed for reasoning-driven retrieval in medicine. It comprises eight datasets spanning diverse clinical scenarios, including medical question answering and diagnostic reasoning. Our experiments reveal that existing retrievers perform poorly on R2MED, with the strongest model reaching only 31.4 nDCG@10. While large reasoning models can provide modest improvements (up to 41.4), a significant performance gap remains. R2MED reveals a fundamental challenge: effective retrieval in medicine requires reasoning, not just semantic matching. In the future, we plan to develop retrieval methods explicitly tailored for reasoning-driven retrieval tasks. Furthermore, we see promising opportunities in extending R2MED to multimodal medical retrieval, incorporating imaging data. Overall, we hope R2MED lays the groundwork for future research into retrieval systems that meet the complex reasoning demands of medical applications.

7 Limitations.

While R2MED covers a diverse set of medical retrieval scenarios and is designed to enable reasoning-centric evaluation, it still has several limitations. First, the benchmark focuses exclusively on text-based retrieval and does not account for multimodal evidence such as medical images, laboratory results, or other structured clinical data, which are often essential in real-world medical decision-making. Second, query filtering, relevance annotation, and parts of the benchmark construction pipeline rely on large language models, whose outputs may be affected by limitations in medical knowledge, instruction-following ability, and model-specific biases.

Although we adopt a two-stage expert review process to improve annotation quality and remove problematic instances, the resulting labels should not be regarded as entirely free of subjectivity or systematic error. Third, the relevant document mining stage depends on a limited set of retrievers and top- k retrieval results, which means some truly relevant documents may be missed and remain unjudged due to model bias or recall limitations, potentially affecting evaluation outcomes. Finally, despite the diversity of sources and query types included in R2MED, the benchmark inevitably reflects source and distribution biases and cannot fully represent the full spectrum of medical information needs across institutions, populations, and specialties. More broadly, R2MED is intended to study reasoning-driven retrieval rather than to capture the complete workflow of clinical decision-making, which in practice involves longitudinal patient context, multimodal evidence, expert interpretation, and downstream action. Residual risks include model-family bias from o3/GPT-4o, reasoning-style bias in generated paths, source/style bias across webpages, Wikipedia, textbooks, PubMed, and IYi cases, and translation artifacts in IYi-Clinical; filtering and expert review mitigate but cannot fully eliminate them.

8 Ethical Considerations

To construct R2MED, we exclusively curated data from publicly available sources, including published medical resources, publicly accessible medical platforms, and anonymized clinician-authored case materials. No identifiable personal health information (PHI) was intentionally collected, used, or disclosed in the process, and we followed estab-

lished data collection procedures to comply with applicable copyright and privacy requirements, including the removal or anonymization of personal identifiers where necessary. Nevertheless, because the benchmark contains medical content and clinical scenarios, it may still include medically sensitive or potentially distressing material. R2MED is developed and released solely for research and academic evaluation in a controlled setting, with the goal of studying reasoning-driven medical retrieval. It is not intended for clinical use, medical diagnosis, treatment planning, or real-world decision-making under any circumstances. Accordingly, outputs from models evaluated on R2MED must be interpreted with caution and should never be used as a substitute for professional medical advice or expert judgment.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. 2016. A full-text learning to rank dataset for medical information retrieval. In *Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20–23, 2016. Proceedings 38*, pages 716–722. Springer.
- burgerbee. 2025. medicine_wiki dataset. https://huggingface.co/datasets/burgerbee/medicine_wiki. Accessed: 2025-03-16.
- Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.
- Jianlv Chen, Nan Wang, Chaofan Li, Bo Wang, Shitao Xiao, Han Xiao, Hao Liao, Defu Lian, and Zheng Liu. 2024b. Air-bench: Automated heterogeneous information retrieval benchmark. *arXiv preprint arXiv:2412.13102*.
- Junyong Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024c. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*.
- MTEB Contributors. 2026. Mteb leaderboard. <https://huggingface.co/spaces/mteb/leaderboard>. Accessed: 2026-03-16.

668	Felix Faltings, Wei Wei, and Yujia Bao. 2025. Enhancing retrieval systems with inference-time logical reasoning . In <i>Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)</i> , pages 449–463, Vienna, Austria. Association for Computational Linguistics.	
669		
670		
671		
672		
673		
674	Yongqi Fan and 1 others. 2025. Medeureka: A medical domain benchmark for multi-granularity and multi-data-type embedding-based retrieval. In <i>Findings of the Association for Computational Linguistics: NAACL 2025</i> .	
675		
676		
677		
678		
679	Zhihao Fan, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, Jun Xi, Fei Huang, and Jingren Zhou. 2024. Ai hospital: Benchmarking large language models in a multi-agent medical interaction simulator. <i>arXiv preprint arXiv:2402.09742</i> .	
680		
681		
682		
683		
684	Giacomo Frisoni, Miki Mizutani, Gianluca Moro, and Lorenzo Valgimigli. 2022. Bioreader: a retrieval-enhanced text-to-text transformer for biomedical literature. In <i>Proceedings of the 2022 conference on empirical methods in natural language processing</i> , pages 5770–5793.	
685		
686		
687		
688		
689		
690	gamino. 2024. <code>wiki_medical_terms</code> dataset. https://huggingface.co/datasets/gamino/wiki_medical_terms . Accessed: 2025-03-16.	
691		
692		
693	Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2022. Precise zero-shot dense retrieval without relevance labels. <i>arXiv preprint arXiv:2212.10496</i> .	
694		
695		
696	Lorraine Goeuriot, Gareth JF Jones, Liadh Kelly, Henning Müller, and Justin Zobel. 2016. Medical information retrieval: introduction to the special issue. <i>Information Retrieval Journal</i> , 19:1–5.	
697		
698		
699		
700	Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. <i>arXiv preprint arXiv:2407.21783</i> .	
701		
702		
703		
704		
705	Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. <i>arXiv preprint arXiv:2501.12948</i> .	
706		
707		
708		
709		
710		
711	Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. <i>arXiv preprint arXiv:2009.03300</i> .	
712		
713		
714		
715	Ruihui Hou, Shencheng Chen, Yongqi Fan, Lifeng Zhu, Jing Sun, Jingping Liu, and Tong Ruan. 2024. Msdiagnosis: An emr-based dataset for clinical multi-step diagnosis. <i>arXiv preprint arXiv:2408.10039</i> .	
716		
717		
718		
719	IiYi. 2026. Iiyi online consultation platform. https://bingli.iiyi.com/ . Accessed: 2026-03-16.	
720		
	Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. <i>arXiv preprint arXiv:2112.09118</i> .	721
		722
		723
		724
		725
	Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. <i>arXiv preprint arXiv:2412.16720</i> .	726
		727
		728
		729
		730
	Mingyi Jia, Junwen Duan, Yan Song, and Jianxin Wang. 2024. medikal: Integrating knowledge graphs as assistants of llms for enhanced clinical diagnosis on emrs. <i>arXiv preprint arXiv:2406.14326</i> .	731
		732
		733
		734
	Shuyang Jiang, Yusheng Liao, Zhe Chen, Ya Zhang, Yanfeng Wang, and Yu Wang. 2025. Meds ³ : Towards medical slow thinking with self-evolved soft dual-sided process supervision. <i>arXiv preprint arXiv:2501.12051</i> .	735
		736
		737
		738
		739
	Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. <i>arXiv preprint arXiv:2503.09516</i> .	740
		741
		742
		743
		744
	Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. <i>Applied Sciences</i> , 11(14):6421.	745
		746
		747
		748
		749
	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. <i>Bioinformatics</i> , 39(11):btad651.	750
		751
		752
		753
		754
		755
	Mohammad Khodadad and 1 others. 2024. Towards domain specification of embedding models in medicine. <i>arXiv preprint</i> .	756
		757
		758
	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. In <i>Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles</i> .	759
		760
		761
		762
		763
		764
		765
	Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. Nv-embed: Improved techniques for training llms as generalist embedding models. <i>arXiv preprint arXiv:2405.17428</i> .	766
		767
		768
		769
		770
	Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. <i>arXiv preprint arXiv:1906.00300</i> .	771
		772
		773
		774

775	Chaofan Li, Zheng Liu, Jianlyv Chen, Defu Lian, and Yingxia Shao. 2025a. Reinforced information retrieval. <i>arXiv preprint arXiv:2502.11562</i> .	open-domain question answering. <i>arXiv preprint arXiv:2009.08553</i> .	831
776			832
777			
778	Lei Li, Xiangxu Zhang, Xiao Zhou, and Zheng Liu. 2024. Automir: Effective zero-shot medical information retrieval without relevance labels. <i>arXiv preprint arXiv:2410.20050</i> .	Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. 2024. Sfembedding-mistral: enhance text retrieval with transfer learning. <i>Salesforce AI Research Blog</i> , 3:6.	833
779			834
780			835
781			836
782	Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025b. Search-o1: Agentic search-enhanced large reasoning models. <i>arXiv preprint arXiv:2501.05366</i> .	Gabriel de Souza P Moreira, Radek Osmulski, Mengyao Xu, Ronay Ak, Benedikt Schifferer, and Even Oldridge. 2024. Nv-retriever: Improving text embedding models with effective hard-negative mining. <i>arXiv preprint arXiv:2407.15831</i> .	837
783			838
784			839
785			840
786			841
787	Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In <i>Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval</i> , pages 2356–2362.	Niklas Muennighoff, SU Hongjin, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2024. Generative representational instruction tuning. In <i>ICLR 2024 Workshop: How Far Are We From AGI</i> .	842
788			843
789			844
790			845
791			846
792			
793			847
794			848
795	Hao Liu, Zhengren Wang, Xi Chen, Zhiyu Li, Feiyu Xiong, Qinhan Yu, and Wentao Zhang. 2025a. HoP-RAG: Multi-hop reasoning for logic-aware retrieval-augmented generation . In <i>Findings of the Association for Computational Linguistics: ACL 2025</i> , pages 1897–1913, Vienna, Austria. Association for Computational Linguistics.	Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. <i>arXiv preprint arXiv:2210.07316</i> .	849
796			
797			850
798			851
799			852
800			853
801			
802	Zheng Liu, Chaofan Li, Shitao Xiao, Chaozhuo Li, Defu Lian, and Yingxia Shao. 2025b. Matryoshka re-ranker: A flexible re-ranking architecture with configurable depth and width. <i>arXiv preprint arXiv:2501.16302</i> .	Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human-generated machine reading comprehension dataset.	854
803			855
804			856
805			
806			857
807	Zheng Liu, Chaofan Li, Shitao Xiao, Chaozhuo Li, Defu Lian, and Yingxia Shao. 2025c. Reasonrank: Teaching student models to rank with reasoning-based knowledge distillation. <i>arXiv preprint arXiv:2508.07050</i> .	Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. <i>arXiv preprint arXiv:2003.06713</i> .	858
808			859
809			
810			860
811			861
812	Gang Luo, Chunqiang Tang, Hao Yang, and Xing Wei. 2008. Medsearch: a specialized search engine for medical information retrieval. In <i>Proceedings of the 17th ACM conference on Information and knowledge management</i> , pages 143–152.	OpenAI. 2024. New embedding models and api updates .	862
813			
814			863
815			864
816			865
817	Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2024. Fine-tuning llama for multi-stage text retrieval. In <i>Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval</i> , pages 2421–2425.	Ankit Pal, Logesh Kumar Umapathi, and Malaikanan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In <i>Conference on health, inference, and learning</i> , pages 248–260. PMLR.	866
818			867
819			868
820			869
821			870
822			871
823	Shengyu Mao, Yong Jiang, Boli Chen, Xiao Li, Peng Wang, Xinyu Wang, Pengjun Xie, Fei Huang, Hua-jun Chen, and Ningyu Zhang. 2024. Rafe: ranking feedback improves query rewriting for rag. <i>arXiv preprint arXiv:2405.14431</i> .	Pengcheng Qiu, Chaoyi Wu, Shuyu Liu, Weike Zhao, Zhuoxia Chen, Hongfei Gu, Chuanjin Peng, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2025. Quantifying the reasoning abilities of llms on real-world clinical cases. <i>arXiv preprint arXiv:2503.04691</i> .	872
824			873
825			
826			874
827			875
828	Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2020. Generation-augmented retrieval for	Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. <i>OpenAI blog</i> , 1(8):9.	876
829			877
830			
		Stephen Robertson, Hugo Zaragoza, and 1 others. 2009. The probabilistic relevance framework: Bm25 and beyond. <i>Foundations and Trends® in Information Retrieval</i> , 3(4):333–389.	878
			879
			880
			881

882	Tao Shen, Guodong Long, Xiubo Geng, Chongyang	Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min	933
883	Tao, Tianyi Zhou, and Daxin Jiang. 2023. Large	Xu, Filippo Menolascina, and Vicente Grau. 2024.	934
884	language models are strong zero-shot retriever. <i>arXiv</i>	Medical graph rag: Towards safe medical large lan-	935
885	<i>preprint arXiv:2304.14233</i> .	guage model via graph retrieval-augmented genera-	936
		tion. <i>arXiv preprint arXiv:2408.04187</i> .	937
886	Stack Exchange Network. 2026. Stack exchange.	Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muen-	938
887	https://stackexchange.com . Accessed: 2026-03-	nighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack:	939
888	16.	Packed resources for general chinese embeddings. In	940
		<i>Proceedings of the 47th international ACM SIGIR</i>	941
889	Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang,	<i>conference on research and development in informa-</i>	942
890	Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A	<i>tion retrieval</i> , pages 641–649.	943
891	Smith, Luke Zettlemoyer, and Tao Yu. 2022. One		
892	embedder, any task: Instruction-finetuned text em-	Yunfei Xie, Juncheng Wu, Haoqin Tu, Siwei Yang,	944
893	beddings. <i>arXiv preprint arXiv:2212.09741</i> .	Bingchen Zhao, Yongshuo Zong, Qiao Jin, Cihang	945
		Xie, and Yuyin Zhou. 2024. A preliminary study of	946
894	Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi,	o1 in medicine: Are we closer to an ai doctor? <i>arXiv</i>	947
895	Niklas Muennighoff, Han-yu Wang, Haisu Liu, Quan	<i>preprint arXiv:2409.15277</i> .	948
896	Shi, Zachary S Siegel, Michael Tang, and 1 others.		
897	2024. Bright: A realistic and challenging bench-	Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong	949
898	mark for reasoning-intensive retrieval. <i>arXiv preprint</i>	Zhang. 2024a. Benchmarking retrieval-augmented	950
899	<i>arXiv:2407.12883</i> .	generation for medicine. In <i>Findings of the Associa-</i>	951
		<i>tion for Computational Linguistics ACL 2024</i> , pages	952
900	Qwen Team. 2024. Qwen2.5: A party of foundation	6233–6251.	953
901	models .		
902	Qwen Team. 2025a. Qwen3: Think deeper, act faster .	Guangzhi Xiong, Qiao Jin, Xiao Wang, Minjia Zhang,	954
		Zhiyong Lu, and Aidong Zhang. 2024b. Improving	955
903	Qwen Team. 2025b. Qwq-32b: Embracing the power	retrieval-augmented generation in medicine with it-	956
904	of reinforcement learning .	erative follow-up questions. In <i>Biocomputing 2025:</i>	957
		<i>Proceedings of the Pacific Symposium</i> , pages 199–	958
905	Nandan Thakur, Nils Reimers, Andreas Rücklé, Ab-	214. World Scientific.	959
906	hishek Srivastava, and Iryna Gurevych. 2021. Beir:		
907	A heterogenous benchmark for zero-shot evalua-	Ran Xu, Wenqi Shi, Yue Yu, Yuchen Zhuang, Yanqiao	960
908	tion of information retrieval models. <i>arXiv preprint</i>	Zhu, May D Wang, Joyce C Ho, Chao Zhang, and	961
909	<i>arXiv:2104.08663</i> .	Carl Yang. 2024. Bmretriever: Tuning large language	962
		models as better biomedical text retrievers. <i>arXiv</i>	963
910	Ellen Voorhees, Tasmee Alam, Steven Bedrick, Dina	<i>preprint arXiv:2404.18443</i> .	964
911	Demner-Fushman, William R Hersh, Kyle Lo, Kirk		
912	Roberts, Ian Soboroff, and Lucy Lu Wang. 2021.	Wen-wai Yim, Asma Ben Abacha, Yajuan Fu, Zhaoyi	965
913	Trec-covid: constructing a pandemic information re-	Sun, Fei Xia, Meliha Yetisgen-Yildiz, and Martin	966
914	trieval test collection. In <i>ACM SIGIR Forum</i> , vol-	Krallinger. 2024. Overview of the mediqa-m3g 2024	967
915	ume 54, pages 1–12. ACM New York, NY, USA.	shared task on multilingual multimodal medical	968
		answer generation. In <i>Proceedings of the 6th Clinical</i>	969
916	Voyage AI. 2024. voyage-3 & voyage-3-lite: A new	<i>Natural Language Processing Workshop</i> , pages 581–	970
917	generation of small yet mighty general-purpose em-	589.	971
918	bedding models .		
919	David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu	Xiangxu Zhang, Lei Li, Yanyun Zhou, Xiao Zhou,	972
920	Wang, Madeleine van Zuylen, Arman Cohan, and	Yingying Zhang, and Xian Wu. 2025. Inflated	973
921	Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying	excellence or true performance? rethinking medi-	974
922	scientific claims. <i>arXiv preprint arXiv:2004.14974</i> .	cal diagnostic benchmarks with dynamic evaluation .	975
		<i>Preprint</i> , arXiv:2510.09275.	976
923	Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang,	Xiangxu Zhang, Xiao Zhou, Hongteng Xu, and Jianxun	977
924	Rangan Majumder, and Furu Wei. 2023a. Improving	Lian. 2026. Hypemed: Enhancing medication recom-	978
925	text embeddings with large language models. <i>arXiv</i>	mendations with hypergraph-based patient relation-	979
926	<i>preprint arXiv:2401.00368</i> .	ships. <i>ACM Transactions on Information Systems</i> .	980
927	Liang Wang, Nan Yang, and Furu Wei. 2023b.	Zhengyun Zhao, Qiao Jin, Fangyuan Chen, Tuorui Peng,	981
928	Query2doc: Query expansion with large language	and Sheng Yu. 2023. A large-scale dataset of patient	982
929	models . In <i>Proceedings of the 2023 Conference on</i>	summaries for retrieval-based clinical decision sup-	983
930	<i>Empirical Methods in Natural Language Processing</i> ,	port systems. <i>Scientific data</i> , 10(1):909.	984
931	pages 9414–9423, Singapore. Association for Com-		
932	putational Linguistics.	Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai	985
		Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and	986
		Bowen Zhou. 2025. Medxpertqa: Benchmarking	987
		expert-level medical reasoning and understanding.	988
		<i>arXiv preprint arXiv:2501.18362</i> .	989

A Dataset Construction

In this section, we provide more detailed information about the four stages of dataset construction. Tables 8-9 summarize the sources of queries and documents for each dataset. The number of queries across different construction stages can be found in Tables 5-7.

A.1 Data Collection

Q&A Reference Retrieval Datasets For the Biology dataset, we directly adopt the version curated by BRIGHT (Su et al., 2024). To broaden domain coverage, we additionally construct two new datasets sourced from the Bioinformatics and Medical Sciences communities on StackExchange. We select posts where the accepted answer has received more than three upvotes and contains at least one external URL. For each selected post, we extract 1-2 linked webpages to serve as the initial positive documents.

To construct the initial negative pool, we use domain-relevant Wikipedia corpora distinct from the sources of positive documents to avoid content overlap. Specifically, we use the *medicine_wiki* corpus (burgerbee, 2025) for the Medical Sciences dataset and the *wiki_medical_terms* (gamino, 2024) for the Bioinformatics dataset. All documents, from external webpages and Wikipedia, are segmented into smaller passages by sentence-level splitting and regrouped into chunks of approximately 128 tokens.

Clinical Evidence Retrieval Datasets The clinical evidence retrieval task comprises three datasets, each representing a critical stage in clinical decision-making: examination recommendation, disease diagnosis, and treatment planning. We construct these datasets based on three representative medical question-answering sources: MedXpertQA (Zuo et al., 2025) (1,861 multiple-choice questions), MedQA (Jin et al., 2021) (1,273 multiple-choice questions), and MedRBench_Treat (Qiu et al., 2025) (496 open-ended questions). To ensure that queries genuinely require medical reasoning, we apply a multi-stage filtering and reformulation pipeline:

- **Task-based Filtering.** We use GPT-4o to annotate each question with a clinical task type (examination, diagnosis, or treatment) and discard those not belonging to the targeted categories. The instruction is shown in Figure 5.

- **Rule-based Filtering.** We further remove questions that do not require reasoning or whose answers are not specific medical entities. This step is automated via GPT-4o with rule-based guidance (see Figure 6).

- **Difficulty Filtering.** To retain only challenging questions, we evaluate each multiple-choice item using four small-scale instruction-tuned models: Qwen2.5-7B-Instruct, Llama3.1-8B-Instruct, Gemma-2-9B-it, and Qwen-14B-Instruct. Questions correctly answered by more than one model are filtered out.

- **Open-ended Reformulation.** Selected multiple-choice questions are reformulated into open-ended formats, with corresponding answers extracted using GPT-4o. The transformation prompt is shown in Figure 7.

For negative corpus construction, we use different sources for each dataset. For the examination dataset (MedXpertQA-Exam), we sample from the Wikipedia subset of the MedCorp corpus (Xiong et al., 2024a). For the diagnosis dataset (MedQA-Diag), we use medical textbook materials released with the original benchmark (Jin et al., 2021). For the treatment dataset (PMC-Treatment), we retain the original article associated with each question as the positive document. To build a challenging negative set, we crawl approximately 14,000 case reports from the PubMed Central Open Access (PMC OA) Subset¹, focusing on case reports tagged with diagnosis or treatment topics.

¹<https://pmc.ncbi.nlm.nih.gov/tools/openftlist/>

Instruction for Task-based Filtering

You are a medical expert. Your task is to classify a given medical Question-Answer (QA) pair into one of the following categories based on the question's intent and the answer's entity type:

1. Examinations Recommendation (EXM): The question asks for the most appropriate diagnostic test or examination to confirm a suspected condition. The answer should be Laboratory Tests, Imaging Examinations, Endoscopic Examinations, or Other Examinations.

2. Diagnostic Reasoning (DIA): The question asks for the most likely disease, syndrome, etiology, or functional disorder affecting the patient. The answer should be Disease Diagnosis, Syndrome Diagnosis, Etiological Diagnosis, or Functional Disorder Diagnosis.

3. Treatment Planning (TRT): The question asks for the best treatment plan, including pharmacological, surgical, or preventive measures. The answer should be Pharmacological Treatment, Surgical Treatment, Other Therapies, and Preventive Measures.

****Task:****

For each given Question-Answer (QA) pair, determine the most appropriate classification from the three categories above. If the QA pair does not fit any category, return "Other".

****Output Format:****

Your output should follow the following format, do not output any additional content:

- Classification: [one of EXM/DIA/TRT/Other]
- Reasoning: [brief explanation of why this classification was chosen]

Here is the Question-Answer (QA) pair:

<Question>: {QUESTION}

<Answer>: {ANSWER}

Figure 5: Instruction for filtering questions based on the task.

Instruction for Rule-based Filtering

You are an expert in filtering and evaluating multiple-choice questions for advanced reasoning tasks. Your task is to evaluate a given question and determine whether it meets the following criteria:

1. Depth of Reasoning: The question should require deeper reasoning. If the question appears too simple, mark it as "Too Simple".

2. Unambiguous Correct Answer: The question must have a unique and unambiguous correct answer. If the question asks for "incorrect options" or allows for multiple correct answers, mark it as "Ambiguous Answer".

3. Open-Ended Reformulation Feasibility: The question should be suitable for reformulating into an open-ended format. If the question cannot be easily reformulated into an open-ended problem and a clear ground-truth answer, mark it as "Not Reformulatable".

4. Medical Entity as the Correct Answer: The correct answer must be a medical entity, such as a disease, drug, symptom, anatomical structure, laboratory test, imaging examination, or treatment method. If the correct option is an abstract concept, behavior, tool, or any non-medical entity, mark it as "Non-Medical Entity".

For each question, provide one of the following evaluations:

- "Pass" (The question meets all the criteria.)
- "Too Simple"
- "Ambiguous Answer"
- "Not Reformulatable"
- "Non-Medical Entity"

****Output Format:****

Your output should follow the following format, do not output any additional content:

- Evaluation: [one of the five evaluations]
- Explanation: [Brief explanation]

Here is the Multiple-choice Question and its answer:

<Question>: {QUESTION} {OPTION}

<Answer>: {ANSWER}

1071

Figure 6: Instruction for filtering questions based on rules.

Instruction for Open-ended Reformulation

You will be given a multiple-choice clinical question. Your task is to transform it into an open-ended question while preserving the original language and structure as much as possible.

Follow these steps:

1. Remove the multiple-choice options from the original question.

2. If the original question contains phrases like "Which of the following...", rewrite it into a self-contained open-ended form, but only minimally modify the wording required to make it complete without the options.

****Output Format:****

Your output should follow the following format, do not output any additional content:

- Open-ended Question: [your rewritten question]
- Standard Answer: [a concise standard answer]

Here is the Multiple-choice Question and its answer:

<Question>: {QUESTION} {OPTION}

<Answer>: {ANSWER}

1072

Figure 7: Instruction for reformatting open-question.

Instruction for Quality Filtering

You are given a medical article containing the following components: Title, Abstract, Case 1 (clinical case), and Case 2 (clinical case). Your task is to assess whether this article is suitable for constructing a similar-case retrieval dataset focused on disease diagnosis.

Evaluate the article using three independent criteria, each scored from 0 to 10, based on the following standards:

Scoring Scale (applies to all three criteria):

- 10: Fully satisfies the criterion
- 7-9: Mostly satisfies the criterion, only minor issues
- 4-6: Partially satisfies the criterion, some key limitations
- 1-3: Minimally satisfies the criterion, major gaps
- 0: Does not satisfy the criterion at all

****Scoring Criteria:****

Score 1 – Diagnostic Focus: Does the title and abstract indicate that the main goal of the article is to study or explain the diagnosis process or mechanism of a specific disease shared across the cases?

Score 2 – Case Completeness & Shared Diagnosis: Do both cases contain sufficient clinical details? Is the same diagnosis clearly stated in both cases?

Score 3 – Diagnostic Similarity & Supportiveness: Are the clinical presentations of the two cases similar? Does Case 2 provide diagnostic insight or evidence that could support the diagnosis in Case 1?

****Output Format:****

Your output should follow the following format, do not output any additional content:

- Score 1: 0-10
- Score 2: 0-10
- Score 3: 0-10
- Explanation: [An overall explanation]

Here is the article:
{ARTICLE}

Figure 8: Instruction for filtering cases based on quality.

Clinical Case Retrieval Datasets For the clinical case retrieval task, we collect patient case records from two primary sources: (1) PMC-Patients (Zhao et al., 2023), a curated collection of multi-case clinical reports from PubMed Central, and (2) IYi-Clinical, a dataset we construct by crawling 10k anonymized patient records from over ten departments on the IYi online consultation platform. All collected records undergo strict de-identification and privacy-preserving processing. We use GPT-4o-mini to translate IYi’s data into the corresponding English version. Dataset construction follows a three-stage pipeline:

- **Task Filtering.** For PMC-Patients, we identify multi-case articles and extract only the first described case in each as the query source. For IYi-Clinical, we group patient records by diagnostic label using rule-based matching. One case is selected as the query, and the remaining cases within the same group constitute the

candidate retrieval pool.

- **Quality Filtering.** Each candidate case is assessed by GPT-4o across three dimensions: diagnostic focus, case completeness, and diagnostic similarity. Only cases that meet predefined thresholds on all three criteria are retained (see Figure 8).
- **Question Formulation.** We use GPT-4o to extract the patient’s clinical presentation from each full case, removing any diagnostic reasoning or outcome information(see Figure 9).

For each query case, we construct the initial positive set by selecting full case records that share the same diagnostic results. All remaining cases in the corpus are treated as initial negatives. In the PMC-Patients dataset, where articles often contain multiple related cases, we retain only 1–3 additional cases from the same report as positive and exclude the remaining ones in the report.

Instruction for Question Formulation

You will be given a clinical case and its corresponding diagnosis (disease entities). Your task is to construct a <Query, Answer> pair in Chinese from the Case, focusing only on disease diagnosis.

Follow these steps:

Step 1. Construct the Query from the Case: Rewrite the Case into a clinical question that simulates a physician seeking diagnostic guidance. Retain detailed clinical findings: symptoms, labs, imaging, history, etc. Remove any explicit diagnostic statements. Ensure the resulting description is clear, coherent, and self-contained. End the query with a natural disease diagnostic question.

Step 2. Write the Answer: Extract the diagnostic reasoning and final diagnosis from the case. The answer should: Contain the reasoning steps or evidence used to reach the diagnosis. End with the confirmed disease entity as the final diagnosis. The disease names should match the given disease entity list.

****Output Format:****

Your output should follow the following format, do not output any additional content:

- Query: Rewritten case 1 with diagnostic question
- Answer: Diagnosis that answers the query

Here are two cases and their disease entities:

<Case>: {CASE}
<Disease>: {DISEASE}

Figure 9: Instruction for rewriting question.

A.2 Relevant Document Mining

For each query, we use OpenAI o3 model to generate a step-by-step reasoning path, following the instructions detailed in Figure 10. This yields a structured triplet <query, reasoning path, answer>, which we refer to as the multi-view retrieval set.

1120 To mine potentially relevant documents, we deploy
 1121 a retrieval committee comprising BM25, MedCPT,
 1122 and BGE-Large, ensuring complementary retrieval
 1123 capabilities. Each element in the triplet is used
 1124 independently as a retrieval query under each re-
 1125 triever, and the top-100 documents are retrieved.

Instruction for Reasoning Path

Please analyze the given medical question and system-
 atically reason through the problem to reach the correct
 answer.
 <Question>:
 {QUESTION}
 <Correct Answer>: {ANSWER}

Figure 10: Instruction for generating reasoning path.

A.3 Relevance Assessment

To evaluate the relevance between each query and
 its corresponding potential positive documents, we
 employ GPT-4o as the assessment model. The
 detailed instruction is illustrated in Figure 11-13.

Instruction for Relevance Assessment (I)

You are an expert in evaluating passages linked in medical
 Q&A forum posts. Each post consists of a Question and a
 long Answer (containing reasoning and conclusions). You
 are given one such passage, and your task is to assess how
 well the passage supports or enhances understanding of the
 post.
 Your evaluation consists of two independent scoring
 aspects, each rated on a 10-point scale:
 Score 1. Coverage of Critical Information (10 points)
 10 points (Highly Relevant): Passage provides detailed
 and accurate information in the post.
 7-9 points (Moderately Relevant): Passage contains
 strong related information but may lack depth.
 4-6 points (Weakly Relevant): Passage vaguely touches
 on ideas in the post but in a general or indirect way.
 0-3 points (Not Relevant): Passage does not meaning-
 fully overlap with the post content or is tangential.
 Score 2. Contribution to the Answer's Reasoning (10
 points)
 10 points (Strong Support): The passage supports at
 least one critical reasoning step with detailed evidence.
 7-9 points (Moderate Support): The passage provides
 useful but incomplete support for a reasoning step.
 4-6 points (Weak Support): The passage only partially
 supports the reasoning process, with limited relevance.
 0-3 points (No Support): The passage does not
 contribute to the reasoning process in any meaningful way.
 Then provide a clear explanation for your score, detailing
 which part(s) of the passage are connected to the answer,
 whether and how it contributes to understanding or
 reasoning.
 Output Format:
 Your output should follow the following format, do not
 output any additional content:
 - Score 1: [Score between 0 and 10 for the Relevance of
 Key Knowledge Points]
 - Score 2: [Score between 0 and 10 for the Support of

the Reasoning Process]
 - Explanation: [Detailed explanation]
 Here is the Post and Passage:
 <Post>: {QUESTION}
 <Passage>: {PASSAGE}

Figure 11: Instruction for relevance assessment on Q&A
 reference retrieval datasets.

Instruction for Relevance Assessment (II)

You are an expert in evaluating retrieved medical passages
 for their relevance and usefulness in supporting clinical
 reasoning. Your task is to assess how well a passage aligns
 with a given question-answer pair, its reasoning process,
 and the key knowledge points.
 Scoring Criteria:
 Score 1. Relevance to Key Knowledge Points (10 points):
 10 points (Highly Relevant): The passage provides
 detailed and precise information directly covering the key
 knowledge points.
 7-9 points (Moderately Relevant): The passage is
 strongly related to the key knowledge but lacks some
 depth.
 4-6 points (Weakly Relevant): The passage touches on
 the knowledge points but lacks detail.
 0-3 points (Not Relevant): The passage does not discuss
 the key knowledge points in a meaningful way or is only
 vaguely related.
 Score 2. Support for the Reasoning Process (10 points):
 10 points (Strong Support): The passage completely
 supports at least one critical reasoning step with detailed
 evidence.
 7-9 points (Moderate Support): The passage provides
 useful but incomplete support for a reasoning step.
 4-6 points (Weak Support): The passage only partially
 supports the reasoning process, with limited relevance.
 0-3 points (No Support): The passage does not
 contribute to the reasoning process in any meaningful way.
 Output Format:
 Your output should follow the following format, do not
 output any additional content:
 - Score 1: [Score between 0 and 10 for the Relevance of
 Key Knowledge Points]
 - Score 2: [Score between 0 and 10 for the Support of
 the Reasoning Process]
 - Explanation: [Detailed explanation]
 Here is the Question-Answer (QA) pair, its reasoning
 process, the key knowledge points, and the medical
 passage:
 <Question>: {QUESTION}, <Answer>: {ANSWER},
 <Reason>: {REASON}
 <Key_point>: {POINT}
 <Passage>: {PASSAGE}

Figure 12: Instruction for relevance assessment on clinical
 evidence retrieval datasets.

Instruction for Relevance Assessment (III)

You are an expert in evaluating retrieved medical case reports for clinical diagnosis support. Your task is to assess how well a case aligns with a target clinical scenario and whether it supports the diagnostic reasoning process leading to the correct diagnosis.

****Scoring Criteria:****

Score 1. Diagnostic Match (10 points):

10 points (Exact Match): The candidate case clearly states the same diagnosis as the answer.

7-9 points (Close Match): The case suggests a clinically similar diagnosis or disease subtypes with the answer.

4-6 points (Partially Related): The case may be related, but the diagnosis is unclear or partially related to the answer.

0-3 points (Unrelated Diagnosis): The diagnosis is clearly different or unrelated.

Score 2. Diagnostic Reasoning Support (10 points):

10 points (Strong Support): The case contains highly similar symptoms, findings, or patterns that mirror key parts of the reasoning process.

7-9 points (Moderate Support): The case overlaps with some reasoning elements but lacks full coverage.

4-6 points (Weak Support): The case only partially relates to the reasoning, with limited diagnostic value.

0-3 points (No Support): The case does not help with diagnostic reasoning in a meaningful way.

****Output Format:****

Your output should follow the following format, do not output any additional content:

- Score 1: [Score between 0 and 10 for Diagnostic Match]

- Score 2: [Score between 0 and 10 for Diagnostic Reasoning Support]

- Explanation: [Detailed explanation]

Here is the Question-Answer (QA) pair, its reasoning process, the candidate case:

<Question>: {QUESTION}, <Answer>: {ANSWER},

<Reason>: {REASON}

<Case>: {CASE}

the reformulated query (if applicable) is self-contained, clinically coherent, and provides sufficient clinical detail for a clinician to answer it. Incomplete or incoherent queries may lack critical patient information or pose ill-formed clinical questions.

- **Plausibility of the Reasoning Path.** This dimension evaluates whether the model-generated reasoning path reflects medically sound logic. High-quality reasoning paths adhere to accepted diagnostic or therapeutic pathways, maintain clinical plausibility, and avoid unsupported or medically invalid inferences.
- **Supportiveness of Positive Documents.** This assesses whether the positive documents provide sufficient and relevant evidence to support the query-answer pair. Strong supporting documents either directly present or clearly imply the necessary findings, differentials, or treatment considerations. Documents lacking topical relevance or clinical substance receive lower ratings.

In total, we review 833 query-answer pairs and approximately 2,500 associated positive documents across the seven datasets. Figure 14 illustrates an example from our annotation platform. The evolution of query counts through different processing stages is summarized in Table 5-7. Fewer than 10% of examples in each dataset are excluded after expert review. Common reasons for exclusion include: (1) exam-style or closed-ended queries such as “which of the following is...”; (2) supporting documents that are only loosely related to the query and lack substantive detail and reasoning paths that contain hallucinations; and (3) reasoning paths that contain factual hallucinations. Importantly, in the last case, if the associated positive documents remain clinically relevant, the example is retained despite imperfections in the reasoning path, as the documents still provide value for evaluating retrieval performance.

A.5 Additional Quality Assessment

To provide an independent estimate of label reliability, we conduct an additional expert assessment on sampled query-document pairs. We sample 200 pairs in total, consisting of 25 pairs from each of the eight subsets. Positive and negative examples are balanced where possible so that the assessment covers both relevance and non-relevance decisions.

Figure 13: Instruction for relevance assessment on clinical case retrieval datasets.

A.4 Expert Review

To further ensure the clinical validity and quality of our benchmark, we conduct a two-stage expert review of all examples. In the first stage, a medically trained annotator (a PhD student) reviews the entire dataset. The annotator receives targeted training prior to annotation, including task-specific guidelines, calibration on example cases, and discussions with clinical experts. In the second stage, a medical expert reviews only the examples flagged as problematic and provides final judgments. Before annotation, we explained the usage of data and obtained consent from annotators. Each example is evaluated based on three criteria:

- **Completeness and Coherence of Reformulated Queries.** This criterion assesses whether

Table 5: Number of queries during different stages on Q&A reference task.

Dataset	Init. Query	AI Assessment	Expert Review
Biology	-	-	103
Bioinformatics	104	83	77
Medical Sciences	105	95	88

We invite three clinical doctoral students majoring in medicine from a QS-ranked Top 100 medical college to serve as medical annotators, and all of them have at least one year of clinical experience. Before annotation, we brief the experts on the task setting. We also explain our research objectives, experimental setup, and data usage in detail, explicitly obtain informed consent from all experts, and provide a remuneration of 100 RMB (\$14) per expert and gained consent of data usage. The sampled data contain no protected health information. The three experts independently annotate whether each document is relevant to the corresponding query under the benchmark definition. While annotating, experts were allowed to use tools such as web search and translators. We then aggregate expert labels by majority vote and compare them with the original qrels. The inter-annotator agreement is high, with Fleiss’ $\kappa = 0.828$, and the original qrels agree with the majority expert label for 88.0% of sampled pairs. These results suggest that the LLM-assisted construction pipeline is substantially corrected by automatic filtering and expert validation.

B Data Examples

In Table 15-22, we show more examples in R2MED.

C Data Diversity Analysis

We use GPT-4o to assign each query to one of twelve pre-defined body systems, based on the prompt shown in Figure 16. Since the three StackExchange-derived datasets focus on general biomedical topics rather than clinical case scenarios, we exclude them from this annotation process and apply the labeling only to the remaining five datasets. In addition, we follow BEIR (Thakur et al., 2021) and compute pairwise weighted Jaccard similarity scores between datasets to evaluate corpus-level distributional diversity. Each corpus is tokenized using the GPT-2 tokenizer, and overlap is measured at the token level. Low inter-dataset similarity confirms that R2MED spans heterogeneous

distributions, presenting a strong generalization challenge for retrieval models. The attribute distributions of R2MED is shown in Figure 15. R2MED covers several reasoning patterns rather than one exhaustive taxonomy: Q&A reference retrieval often requires conceptual or causal explanation, clinical evidence retrieval emphasizes diagnostic and therapeutic inference, and clinical case retrieval relies on abductive matching from observed symptoms to latent diagnoses. We therefore interpret the benchmark as a broad reasoning-driven retrieval testbed, while fine-grained reasoning-type annotation is left for future releases. Cross-dataset performance differences likely reflect both corpus composition and task demands: StackExchange subsets contain heterogeneous web evidence, MedXpertQA-Exam requires specialist exam-style inference, PMC-Treatment has more directly aligned treatment evidence, and IYi-Clinical contains shorter translated consultation records with noisier symptom descriptions.

Instruction for Body System Annotation

You are an experienced medical doctor and independent practitioner. Your task will be to label a medical question according to the human body system it corresponds to. You will be given a list of human body systems, followed by a medical question. Please determine which system the question best pertains to. If the question is related to multiple systems, only select the most relevant one. Directly output the name of the final system you selected from the list of available systems, do not output any additional content.

****Systems:****
Integumentary, Skeletal, Muscular, Nervous, Endocrine, Cardiovascular, Lymphatic, Respiratory, Digestive, Urinary, Reproductive, Immune, Other

****Question:****
{QUESTION}

Figure 16: Instruction for body system annotation.

D Dataset License and Usage

D.1 Dataset License

Table 8-9 summarize the data sources and corresponding licenses for the eight datasets included in R2MED. Most datasets are distributed under permissive licenses, such as variants of the Creative Commons Attribution (CC-BY) license and the MIT license, allowing sharing and adaptation for academic and research purposes. Although the MedRBench (PMC-Treatment) dataset does not explicitly specify a license in its repository, it is derived from PubMed Central Open Access

Raw Data: A 19-year-old Caucasian male presents to your office with hypopigmented skin. He undergoes a skin biopsy and is found to have an absence of melanocytes in the epidermis. Which of the following is the most likely diagnosis? Options: (A) Tinea versicolor (B) Albinism (C) Vitiligo (D) Melanoma. Answer: (C) Vitiligo

<p>Reformatted Query: A 19-year-old Caucasian male presents to your office with hypopigmented skin. He undergoes a skin biopsy and is found to have an absence of melanocytes in the epidermis. What is the most likely diagnosis?</p>	<p>Reasoning Path: Step 1: Identify the key clinical findings. The patient is a 19-year-old Caucasian male. He presents with hypopigmented (depigmented) areas on his skin. A skin biopsy reveals an absence of melanocytes in the epidermis. Step 2: Consider the differential diagnosis for hypopigmentation. Vitiligo: Autoimmune destruction of melanocytes leading to depigmented patches. Pityriasis alba: Generally shows mild hypopigmentation with retained melanocytes. Tinea versicolor: A fungal infection that causes discoloration but does not eliminate melanocytes. Nevus depigmentosus: A</p>	<p>Passage 1: These spots then expand, sometimes in dermatomal patterns, to include large segments of skin. The disease is usually chronic and progressive, with some patients becoming completely depigmented. Many patients have serologic markers of autoimmune disease (e.g., antithyroid antibodies, DM, pernicious anemia) but seldom present with these diseases. Patients with malignant melanoma may develop an antimelanocyte immune response that leads to vitiligo. Diagnosed by the history and clinical picture, with histology demonstrating total absence of melanocytes. Conditions to rule</p>	<p>Passage 2: Absence of melanocytes Decreased number or absence of melanocytes Amelanotic areas—few to no melanocytes Possible somatic mutations as a reflection of aging or UV exposure Block in transfer of melanin from melanocytes to keratinocytes could be secondary to edema or decrease in contact time Destruction of melanocytes if inflammatory cells attack basal layer of epidermis Invasion of stratum corneum by the yeast Yeast is lipophilic and produces C9 and C11 dicarboxylic acids, which in vitro inhibit tyrosinase Autoimmune phenomenon that results in destruction of melanocytes—primarily</p>
<p>Reformatted Query Completeness?</p> <p><input checked="" type="radio"/> 1 <input type="radio"/> 0</p>	<p>Reasoning Path Correctness?</p> <p><input checked="" type="radio"/> 1 <input type="radio"/> 0</p>	<p>Passage 1 Relevance?</p> <p><input checked="" type="radio"/> 1 <input type="radio"/> 0</p>	<p>Passage 2 Relevance?</p> <p><input type="radio"/> 1 <input checked="" type="radio"/> 0</p>
<p>Previous Page</p>		<p>Next Page</p>	
		<p>Jump Current Page</p>	
<p>Submit</p>			
<p>Return to Homepage</p>			


Use via API  · Built with Gradio 

Figure 14: Annotation interface of R2MED.

Table 6: Number of queries during different stages on clinical evidence retrieval task.

Dataset	Init. Query	Task Filter	Rule Filter	Difficulty Filter	Reformatting	AI Assessment	Expert Review
MedXpertQA-Exam	1861	302	220	181	125	105	97
MedQA-Diag	1273	656	543	149	126	124	118
PMC-Treatment	496	496	496	243	243	162	150

Table 7: Number of queries during different stages on clinical case retrieval task.

Dataset	Init. Query	Task Filter	Quality Filter	Rewriting	AI Assessment	Expert Review
PMC-Clinical	167k	2817	158	158	123	114
IYi-Clinical	10k	553	148	148	141	129

(PMC OA) Subset ², which is a publicly available resource widely used in academic research. For IYi-Clinical, the data originates from a publicly accessible medical consultation platform. Prior studies (Yim et al., 2024; Fan et al., 2024; Hou et al., 2024; Jia et al., 2024) confirm that data from this platform, once anonymized, is permissible for research and educational use. In summary, all datasets used in R2MED have been verified to be legally suitable for research.

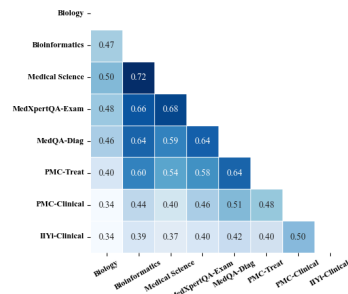
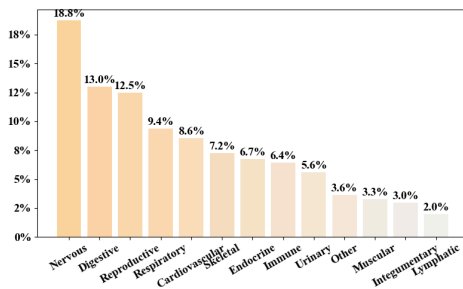
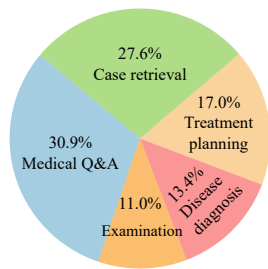
²<https://pmc.ncbi.nlm.nih.gov/tools/openftlist/>

D.2 Dataset Instance Metadata

The codes of R2MED dataset is organized into three files: `query.jsonl`, `corpus.jsonl`, and `qrels.jsonl`, corresponding to queries, corpus passages, and relevance labels, respectively. Each file follows a line-delimited JSON (.jsonl) format. The schema for each file is summarized below:

query.jsonl Each row represents a query:

- `id`: The unique identifier of the query. 1294
- `text`: The textual content of the query. 1295
- `answer`: The intermediate reasoning answer associated with the query. 1296



(a) Distribution of Medical Scenario Types

(b) Distribution of Body Systems

(c) Pairwise similarity scores between datasets

Figure 15: Attribute distributions of R2MED showcase its diversity and comprehensiveness.

- `doc_id`: A list of golden positive document IDs that are relevant to the query.

- `body_system`: The body system category associated with the query.

corpus.jsonl Each row represents a document:

- `id`: The unique identifier of the document.
- `text`: The textual content of the document.

qrels.jsonl Each row represents a query-passage relevance label:

- `q_id`: The ID of the query.
- `p_id`: The ID of the document.
- `score`: The binary relevance score.

Although binary qrels remain the main evaluation protocol, the construction pipeline also records the two-dimensional 0–10 relevance and reasoning-support scores before thresholding. We will release these graded scores as supplementary metadata so future work can study graded relevance, softer losses, and alternative metrics. For the web-based Q&A subsets, evaluation uses the preprocessed static `corpus.jsonl` files rather than live webpages, so later web-content drift does not affect benchmark reproducibility.

D.3 Author Statement

We affirm that all datasets incorporated into R2MED have been verified to originate from sources with open-source or permissive licenses (e.g., CC-BY, MIT). Nonetheless, we fully acknowledge the importance of respecting the rights and concerns of original data providers. Should any licensing issues be identified or brought to our attention, we are committed to responding promptly

Table 8: The license of query sources in R2MED.

Dataset	Source License
Biology	CC-BY 4.0 license
Bioinformatics	CC-BY-SA license
Medical Sciences	CC-BY-SA license
MedXpertQA-Exam	MIT license
MedQA-Diag	MIT license
PMC-Treatment	-
PMC-Clinical	CC-BY-NC-SA 4.0 license
IYi-Clinical	-

Table 9: The license of document sources in R2MED.

Dataset	Source License
Biology	CC-BY 4.0 license
Bioinformatics	GPL 3.0 license
Medical Sciences	CC-BY license
MedXpertQA-Exam	Public Domain (US Government Work)
MedQA-Diag	MIT license
PMC-Treatment	CC BY license
PMC-Clinical	CC-BY-NC-SA 4.0 license
IYi-Clinical	-

and taking appropriate corrective actions. To ensure transparency and traceability, we will maintain versioned releases of R2MED on both HuggingFace and GitHub, and ensure that any future updates (e.g., correction of metadata, expansion of sources).

E Experiment Details

E.1 Benchmark Positioning

R2MED is complementary to recent medical retrieval and embedding benchmarks but differs in how relevance is defined. CMIRB/AutoMIR (Li et al., 2024) focuses on Chinese medical information retrieval without relevance labels, MedEureka (Fan et al., 2025) evaluates medical

embedding retrieval across multiple granularities and data types, and MedTEB (Khodadad et al., 2024) provides a broad medical text-embedding benchmark covering tasks such as classification, clustering, pair classification, and retrieval. In contrast, R2MED targets retrieval where a document is relevant because it supports a latent answer, diagnosis, or treatment plan that is not necessarily visible in the query surface form. This makes R2MED a reasoning-centric benchmark rather than a replacement for broader medical embedding evaluations.

E.2 Model Details

We summarize all retrieval and reranking models used in this study in Table 10, including model names, parameter sizes, and implementation sources. The BM25 baseline is implemented using Pyserini (Lin et al., 2021). For details regarding the large language models and large reasoning models evaluated throughout the paper, please refer to Table 11.

In this work, we evaluate three generation-augmented retrieval (GAR) methods. HyDE (Gao et al., 2022) prompts an instruction-following LLM in a zero-shot setting to generate a hypothetical answer document, which is then used to retrieve relevant information. Query2doc (Wang et al., 2023b) adopts a few-shot prompting strategy to generate pseudo-documents from the query using an LLM and expands the query with these documents to improve retrieval performance. LameR (Shen et al., 2023) augments queries by incorporating potential in-domain answers and prompting an LLM to rewrite the query in a retrieval-friendly form. For search-enhanced large reasoning models, we explore two recent approaches. Search-R1 (Jin et al., 2025) extends DeepSeek-R1 by employing reinforcement learning to enable the model to autonomously generate multiple search queries and retrieve external evidence during multi-step reasoning. In contrast, Search-o1 (Li et al., 2025b) introduces an agent-based retrieval-augmented reasoning framework, incorporating a reason-in-documents module that iteratively refines the evidence selection throughout the reasoning process.

E.3 Evaluation Settings and Instructions

We outline the evaluation instructions used for InstructOR-L, InstructOR-XL, BGE-Large, BMRetriever-410M/2B/7B, E5-mistral, SFR-Embedding, NV-Embed-v2, and GritLM-7B in Table 12. For the embedding model provided by

Table 10: Detailed information on all of the retrieval and reranking models in our paper.

Model	Size	Architecture
Retrieval Models		
BM25 (Robertson et al., 2009)	N/A	Sparse
Contriever (Izacard et al., 2021)	110M	Encoder
MedCPT (Jin et al., 2023)	220M	Encoder
InstructOR-L (Su et al., 2022)	335M	Encoder
BGE-Large (Xiao et al., 2024)	335M	Encoder
BMRetriever (Xu et al., 2024)	410M	Encoder
InstructOR-XL (Su et al., 2022)	1.5B	Encoder
BMRetriever-2B (Xu et al., 2024)	2B	Decoder
E5-mistral (Wang et al., 2023a)	7B	Decoder
BMRetriever-7B (Xu et al., 2024)	7B	Decoder
SFR-Embedding (Meng et al., 2024)	7B	Decoder
GritLM-7B (Muennighoff et al., 2024)	7B	Decoder
NV-Embed-v2 (Lee et al., 2024)	7B	Decoder
Voyage-3 (Voyage AI)	N/A	Dense
OpenAI-3-large (OpenAI)	N/A	Dense
Reranking Models		
MonoBERT (Nogueira et al., 2019)	335M	Encoder
BGE-Reranker (Chen et al., 2024a)	568M	Encoder
RankLLaMA (Ma et al., 2024)	7B	Decoder

Table 11: All LLMs and LRMs used in experiments.

Model	Size
Large Language Models	
Qwen2.5-7B-Ins. (Team, 2024)	7B
Qwen2.5-32B-Ins. (Team, 2024)	32B
Qwen2.5-72B-Ins. (Team, 2024)	72B
Llama3.1-70B-Ins. (Grattafiori et al., 2024)	70B
GPT-4o (Achiam et al., 2023)	N/A
Large Reasoning Models	
R1-Distill-Qwen-32B (Guo et al., 2025)	32B
QwQ-32B (Team, 2025b)	32B
Qwen3-32B (Team, 2025a)	32B
R1-Distill-Llama-70B (Guo et al., 2025)	70B
HuatuoGPT-o1-70B (Chen et al., 2024c)	70B
o3-mini (openai, 2025)	N/A

Voyage, we specify the "input_type" parameter as either "query" or "document" to distinguish queries from documents.

In our experiments, the large language model generates one hypothetical document per query for HyDE. For Query2doc, we manually select two additional in-domain examples from each dataset to construct few-shot prompts as contextual guidance. LameR first retrieves the top-10 documents using BM25. These retrieved documents are then incorporated into the prompt as context to enhance

1405 pseudo-document generation quality. The exact
1406 prompts for all three GAR methods are provided in
1407 Table 13.

1408 For search-enhanced reasoning methods, we follow
1409 the official evaluation settings of Search-R1 (Jin
1410 et al., 2025) and Search-o1 (Li et al., 2025b). Both
1411 systems operate on MedCorp (Xiong et al., 2024a),
1412 a large-scale medical corpus containing 53.9 mil-
1413 lion document chunks. We set the maximum num-
1414 ber of retrieval rounds to 10. In each round, the
1415 system retrieves the top-5 documents, which are
1416 passed into the reasoning model to iteratively syn-
1417 thesize evidence and generate final answers. All
1418 reported results are based on a single run. For
1419 embedding-based retrievers, the retrieval process is
1420 deterministic given fixed queries, documents, and
1421 model checkpoints, and thus does not involve mean-
1422 ingful run-to-run randomness. For LLM-based gen-
1423 eration components (e.g., HyDE, Query2doc, and
1424 reasoning-based methods), we use low-temperature
1425 decoding to minimize stochastic variation and re-
1426 port the results from one run for each setting. Med-
1427 Corp is used only during the search-enhanced rea-
1428 soning and query-rewriting stage. It is not merged
1429 into the R2MED benchmark corpus; after query
1430 rewriting, all final retrieval scores are computed
1431 against the same released R2MED corpus and
1432 qrels.

1433 E.4 Bootstrap Significance Analysis

1434 We additionally conduct a query-level paired boot-
1435 strap test for a representative LRM comparison.
1436 Specifically, we compare NV-Embed-v2 without
1437 LRM rewriting against QwQ-32B under the HyDE
1438 setup using the same NV-Embed-v2 retriever back-
1439 end. For each system, we first compute per-query
1440 nDCG@10 on the R2MED evaluation set. We then
1441 repeatedly resample queries with replacement and
1442 recompute the average paired difference between
1443 the two systems, which yields an empirical con-
1444 fidence interval for the improvement. The repre-
1445 sentative QwQ-32B setting improves nDCG@10
1446 from 31.42 to 36.80, corresponding to an absolute
1447 gain of 5.38 points. The 95% bootstrap confidence
1448 interval is [4.16, 6.56], and the paired bootstrap test
1449 gives $p < 0.001$. This analysis supports that the
1450 observed improvement is unlikely to be explained
1451 by query-sampling variation alone, while the main
1452 paper reports it as representative evidence rather
1453 than an exhaustive significance analysis for every
1454 model pair.

1455 E.5 Computing Resources

1456 All experiments were conducted on a machine with
1457 4 NVIDIA A100 GPUs (40GB each). BM25 was
1458 evaluated on CPU, while all other retrieval models
1459 utilized GPU resources. Evaluation time varied
1460 according to model scale and complexity. For the
1461 retrievers presented in this paper, end-to-end evalu-
1462 ation of a single model generally requires no more
1463 than 8 hours using the 4 A100 GPUs. For methods
1464 involving large language model generation, such
1465 as pseudo-document generation or reasoning, we
1466 leverage vLLM (Kwon et al., 2023) to accelerate
1467 the inference process and reduce latency.

1468 F More Experiment Results

1469 This section provides comprehensive evaluation
1470 results on the R2MED benchmark. Table 23 and
1471 Table 24 report the precision@10 and recall@10
1472 scores of 15 retrieval models. We further present
1473 the performance of generation-augmented retrieval
1474 (GAR) methods in Table 26-28, based on three
1475 underlying retrievers: BM25, BGE-Large, and
1476 NV-Embed-v2. Additionally, Table 29-31 sum-
1477 marize the results of large reasoning models when
1478 combined with BM25, BGE-Large, and OpenAI-3-
1479 large as the retrieval backends.

1480 G Reasoning-Oriented Reranker Results

1481 Table 14 presents externally reported reasoning-
1482 oriented reranker results on R2MED from Reason-
1483 Rank (Liu et al., 2025c). These results are reported
1484 under their own reranking setting and are included
1485 only as complementary evidence, not as part of our
1486 unified main evaluation. In particular, all models
1487 rerank top-100 passages retrieved by E5-mistral,
1488 whereas our main reranking analysis uses the uni-
1489 fied setting described in Section 4.

1490 H Generative Assistance

1491 We employ AI assistants to refine our paper’s con-
1492 tent, and we leverage AI tools to aid in code devel-
1493 opment.
1494

Table 12: Instructions used for benchmarking different datasets for retrieval models.

Instruction Templates for Different Retrievers	
Retriever	Instruction
InstructOR-L, InstructOR-XL	Query: Represent the {tag ₁ } for retrieving relevant paragraphs: Doc: Represent the {tag ₁ } paragraph for retrieval:
BGE-Large	Query: Represent this sentence for searching relevant passages:
BMRetriever-410M/2B/7B	Query: Given a {tag ₁ }, retrieve relevant passages that help answer the {tag ₂ }\n Query: Doc: Represent this passage\n passage:
E5-mistral, SFR-Embedding, NV-Embed-v2	Query: Instruct: Given a {tag ₁ }, retrieve relevant passages that help answer the {tag ₂ }\n Query:
GritLM-7B	Query: < user >\n Given a {tag ₁ }, retrieve relevant passages that help answer the {tag ₁ }< embed >\n Doc: < embed >\n
Tags for Different Datasets	
Dataset	Tag
Biology	tag ₁ : Biology Post, tag ₂ : post
Bioinformatics	tag ₁ : Bioinformatics Post, tag ₂ : post
Medical Sciences	tag ₁ : Medical Sciences Post, tag ₂ : post
MedXpertQA-Exam	tag ₁ : Medical Exam, tag ₂ : exam
MedQA-Diag	tag ₁ : Medical Exam, tag ₂ : exam
PMC-Treatment	tag ₁ : Clinical Case, tag ₂ : case
PMC-Clinical	tag ₁ : Clinical Case, tag ₂ : case
IIYi-Clinical	tag ₁ : Clinical Case, tag ₂ : case

Table 13: Instructions used for evaluating different datasets for generation-augmented retrieval (GAR) methods. {TEXT}, {EXAMPLE}, and {PASSAGE} are the corresponding placeholder.

Method	Instruction
Biology, Bioinformatics, and Medical Sciences datasets	
HyDE	Please write a passage to help answer the {domain} post.\n Post: {TEXT}\n Passage:
Query2Doc	Write a passage that help answer the {domain} post.\n Examples: {EXAMPLE}\n Real Test:\n Post: {TEXT}\n Passage:
LameR	Give a {domain} post and its possible relevant passages (most of these passages are wrong). Please write a correct passage that help answer the post.\n Post: {TEXT}\n Possible Relevant Passages: {PASSAGE}
MedXpertQA-Exam, MedQA-Diag, and PMC-Treatment datasets	
HyDE	Please write a passage to help answer the Medical Exam.\n Exam: {TEXT}\n Passage:
Query2Doc	Write a passage to help answer the Medical Exam.\n Examples: {EXAMPLE}\n Real Test:\n Exam: {TEXT}\n Passage:
LameR	Give a Medical Exam and its possible relevant passages (most of these passages are wrong). Please write a correct passage that help answer the exam.\n Exam: {TEXT}\n Possible Relevant Passages: {PASSAGE}
PMC-Clinical and IYi-Clinical datasets	
HyDE	Please write a similar case to help diagnose the Clinical Case.\n Case: {TEXT}\n Similar Case:
Query2Doc	Write a similar case to help diagnose the Clinical Case.\n Examples: {EXAMPLE}\n Real Test:\n Case: {TEXT}\n Similar Case:
LameR	Give a Clinical Case and its possible similar cases (most of these cases are non-similar). Please write a correct similar case that help diagnose the case.\n Case: {TEXT}\n Possible Similar Cases: {PASSAGE}

Model	Biology	Bioin.	MedS.	MedE.	MedD.	PMCT.	PMCC.	IYiC.	Avg.
Original retriever									
E5-mistral (7B)	18.28	41.47	41.01	6.44	11.38	19.81	30.97	21.37	23.84
Non-reasoning reranker									
RankT5 (3B)	13.20	32.84	23.38	2.09	4.18	0.61	14.55	12.39	12.91
RankZephyr (7B)	22.86	43.07	48.24	6.97	10.45	26.64	7.78	14.59	22.58
Reasoning-oriented reranker									
Rank-R1 (7B)	34.04	51.63	50.96	12.79	21.96	34.79	31.70	25.08	32.87
Rank-R1 (14B)	38.78	53.81	57.94	15.15	25.49	40.57	42.52	29.63	37.99
Rank1 (7B)	32.64	55.57	54.74	12.78	19.98	34.40	30.17	18.15	32.30
Rank1 (32B)	31.84	61.65	59.74	16.56	26.90	41.31	45.56	29.49	39.13
Rearank (7B)	38.35	50.91	59.73	14.06	19.34	37.49	34.29	26.27	35.06
Rank-K (32B)	32.94	53.94	51.47	11.41	22.93	34.44	38.38	26.24	33.97
ReasonRank (7B)	46.80	59.70	60.11	16.48	24.92	39.21	39.13	29.85	39.53
ReasonRank (32B)	45.56	67.73	63.45	18.90	30.60	41.08	46.11	29.35	42.85

Table 14: Externally reported reasoning-oriented reranker results on R2MED, reproduced from Table 2 of ReasonRank (Liu et al., 2025c). All models rerank top-100 passages retrieved by E5-mistral under the external setting; these results are complementary evidence and are not part of our unified main evaluation.

Table 15: An example from Biology dataset.

Query (ID: 225)
<p>How does a tree trunk sprout and grow after being cut? After I cut trees into logs and remove the branches in winter, they start growing. They sprout out and grow completely normal looking stems and leaves and maintain them all summer. The sprouts mostly appear around the cut branches. Sometimes they last all winter and grow for another year. How does it find the energy and water necessary to maintain and grow these stems without ground connection and a water source?</p>
Answer
<p>This is basically the same that happens after pruning (http://en.wikipedia.org/wiki/Pruning) and involves a basic hormonal regulation mechanism in the plants. What happens is that the cut piece of the wood forms a new meristem (http://en.wikipedia.org/wiki/Meristem) which allows the growth of new organs. What's important is that there is no other growth happening nearby, since that would hormonally inhibit any further growth. This is why such growths happen once you've cut the wood, not before (on the healthy stem). This inhibitory effect is known as apical dominance (http://en.wikipedia.org/wiki/Apical_dominance), which has now been disabled. As to where the energy and water comes from, to some extent it is stored within the branches themselves. That's why you need to dry them before being able to use them in a fire. However, this growth is pretty limited. Further water is probably collected by condensation of water vapour in the air.</p>
Positive document (ID: tree_grow_cut/Meristem_1.txt)
<p>Primary meristems[edit] Apical meristems give rise to the primary plant body and are responsible for primary growth, or an increase in length or height. Apical meristems may differentiate into three kinds of primary meristem: Protoderm: lies around the outside of the stem and develops into the epidermis. Procambium: lies just inside of the protoderm and develops into primary xylem and primary phloem. It also produces the vascular cambium, and cork cambium, secondary meristems. The cork cambium further differentiates into the phelloderm (to the inside) and the phellem, or cork (to the outside). All three of these layers (cork cambium, phellem, and phelloderm) constitute the periderm. In roots, the procambium can also give rise to the pericycle, which produces lateral roots in eudicots. Ground meristem: Composed of parenchyma, collenchyma and sclerenchyma cells that develop into the cortex and the pith.</p>
Negative document (ID: tree_grow_cut/how-to-manage-tree-suckers_2_0.txt)
<p>During the summer, your trees will bloom with beautiful leaves in the heat. However, hotter weather usually encourages some tree suckers. To maintain your tree's aesthetic, you'll want to stop these tree suckers from growing. Unfortunately, any tree can suffer from developing the growth of tree suckers, which is likely to happen if the tree is injured or under stress! Not to worry if you've never dealt with these suckers, as stopping tree suckers is simple and effective with the right tools and steps!</p>

Table 16: An example from Bioinformatics dataset.

Query (ID: 21)
<p>How to select high quality structures from the Protein Data Bank?</p> <p>Models of structures deposited in the Protein Data Bank vary in the quality, depending both on the data quality and expertise and patience of the person who built the model. Is there a well-accepted subset of the PDB entries that has only "high quality" structures?</p> <p>Ideally these structures would be representative for classes of proteins in the whole PDB.</p>
Answer
<p>There is a very nice database, pdbcull (http://dunbrack.fccc.edu/Guoli/pisces_download.php) (also known as the PISCES server in the literature). It filters the PDB for high resolution and reduced sequence identity. It also seems to be updated regularly. Depending on the cut-offs, you get between 3000 and 35000 structures.</p> <p>If you are specifically interested in rotamers, you may want to look at top8000 instead, where they have checked for high resolution, and good MolProbity scores. They also provide a rotamer database.</p> <p>PDB also provides their own clustering. They first cluster the sequences, and then extract a representative structure for each one, based on the quality factor ($1/\text{resolution} - R_value$). This has the advantage of being comprehensive, but you will have bad structures when no good ones were ever obtained.</p>
Positive document (ID: Pisces: A_Protein_Sequence_1)
<p>For each calculated list, the server provides an output list of accession IDs (e.g., 1ABCA) with sequence length, structure determination method, resolution, and R-factor (if available) and a file of the sequences in FASTA format. The email containing links to these files will be emailed to the user upon completion of the calculation, and will be stored for at least one week.</p> <p>PDB sequences are updated weekly from the PDB mmCIF files.</p> <p>PISCES correctly handles multi-character chain IDs, which are now used in very large structures by the PDB (and some small structures for no good reason).</p> <p>PISCES now allows the user to select whether to include X-ray, NMR, or cryo-EM structures in the output lists.</p>
Negative document (ID: Amyloid_13)
<p>Combined, these methods have provided 3D atomic structures of amyloid fibrils formed by amyloid β peptides, α-synuclein, tau, and the FUS protein, associated with various neurodegenerative diseases.[46,47] X-ray diffraction studies of microcrystals revealed atomistic details of core region of amyloid, although only for simplified peptides having a length remarkably shorter than that of peptides or proteins involved in disease.[48,49] The crystallographic structures show that short stretches from amyloid-prone regions of amyloidogenic proteins run perpendicular to the filament axis, consistent with the "cross-β" feature of amyloid structure.</p>

Table 17: An example from Medical Sciences dataset.

Query (ID: 2)
<p>Why do corticosteroids harm COVID-19 patients?</p> <p>This Science Daily article states that steroids could do more harm than good in treating coronavirus (COVID-19), referring to this The Lancet article.</p> <p>Understanding the evidence for harm or benefit from corticosteroids in 2019-nCoV is of immediate clinical importance. It's unclear to me how exactly the use of corticosteroids (which is a common treatments for Asthma patients) can do harm other than the vague explanation of steroids also impair the immune system's ability to fight viruses:</p> <p>How exactly does the use of steroids harm the body in case of fighting a virus such as SARS-CoV-2 (2019-nCoV) that causes COVID-19?</p>
Answer
<p>One of the primary purposes of corticosteroids is to suppress immune activity and inflammation: that's exactly why they are used in asthma.</p> <p>Of course the immune system has an actual job besides causing nuisance inflammation: fighting infection.</p> <p>For some infections, the harm to the infected person caused by the immune reaction itself is worse than that of the pathogen itself, so steroids can help prevent damage or ease symptoms while the immune system continues to mount a response, without being hyperactive. For others, immune suppression may be detrimental.</p> <p>Corticosteroids mediate the immune response by suppressing a variety of cytokines and increasing others, which influences activity circulating numbers of different immune cell populations. Cortisol binds the glucocorticoid receptor (https://en.wikipedia.org/wiki/Glucocorticoid_receptor), causing numerous changes in gene expression, many of which are associated with the immune system. From Wikipedia (https://en.wikipedia.org/wiki/Cortisol#Immune_response)</p> <p>In other words, corticosteroids work by by suppressing the generalized killing/cleanup part of the immune system mediated by macrophages and CD8+ cells, while preserving the B-cell part of the immune system that produces specific antibodies and the neutrophils that phagocytose antibody- and complement-bound pathogens.</p> <p>...</p>
Positive document (ID: Cortisol_7)
<p>Immune response[edit] Cortisol prevents the release of substances in the body that cause inflammation. It is used to treat conditions resulting from overactivity of the B-cell-mediated antibody response. Examples include inflammatory and rheumatoid diseases, as well as allergies. Low-dose topical hydrocortisone, available as a nonprescription medicine in some countries, is used to treat skin problems such as rashes and eczema. Cortisol inhibits production of interleukin 12 (IL-12), interferon gamma (IFN-gamma), IFN-alpha, and tumor necrosis factor alpha (TNF-alpha) by antigen-presenting cells (APCs) and T helper cells (Th1 cells), but upregulates interleukin 4, interleukin 10, and interleukin 13 by Th2 cells. This results in a shift toward a Th2 immune response rather than general immunosuppression.</p>
Negative document (ID: Stress_Hormone_6)
<p>However, the organism makes antibodies against this viral protein, and those antibodies also kill the human ACTH hormone, which leads to the suppression of adrenal gland function. Such adrenal suppression is a way for a virus to evade immune detection and elimination.[15][4][6] This viral strategy can have severe consequences for the host (human that is infected by the virus), as cortisol is essential for regulating various physiological processes, such as metabolism, blood pressure, inflammation, and immune response.</p>

Table 18: An example from MedXpertQA-Diag dataset.

Query (ID: Text-162)
An 82-year-old woman with a history of moderate aortic stenosis and a drug-eluting stent for coronary artery disease is admitted to the intensive care unit with a one-day history of symptomatic hypotension and melena. She takes aspirin, clopidogrel, simvastatin, metoprolol, and lisinopril. She is resuscitated with intravenous fluids and blood products, is started on high-dose proton-pump-inhibitor therapy, and stabilizes without further bleeding. Upper and lower endoscopy are unremarkable. What test is the most appropriate next step in evaluating this patient?
Answer
Video-capsule endoscopy
Reasoning path
<p>Let's break down the problem step-by-step:</p> <ol style="list-style-type: none"> 1. Patient Profile and Presentation: The patient is an 82-year-old woman with a history of moderate aortic stenosis and coronary artery disease (with a drug-eluting stent), currently on dual antiplatelet therapy (aspirin and clopidogrel). She presents with symptomatic hypotension and melena, which are suggestive of gastrointestinal bleeding. 2. Initial Management: She was resuscitated with intravenous fluids and blood products. High-dose proton-pump inhibitors were started, and her bleeding stabilized. Upper and lower endoscopies were performed but did not reveal any bleeding source. 3. Diagnostic Challenge: When both upper and lower endoscopies are unremarkable in a patient with suspected GI bleeding, the most likely source is the small bowel. The small bowel is not accessible via standard endoscopy, which is why further evaluation is necessary. 4. Next Diagnostic Step: The next appropriate test for evaluating small bowel bleeding is video-capsule endoscopy. This non-invasive procedure allows visualization of the entire small intestine and can detect lesions like angioectasia, tumors, or other small-bowel abnormalities that could be causing occult bleeding. 5. Conclusion: In summary, once upper and lower endoscopies are negative in a patient with GI bleeding, video-capsule endoscopy is the next best step in evaluation. <p>Final Answer: Video-capsule endoscopy.</p>
Positive document (ID: wiki20220301en160_42780)
Common reasons for using capsule endoscopy include diagnosis of unexplained bleeding, iron deficiency, or abdominal pain, searching for polyps, ulcers and tumors of small intestine, and diagnosis of inflammatory bowel disease. The images collected by the miniature camera during a session are transferred wirelessly to an external receiver worn by the patient, using any one of a band of appropriate frequencies. The collected images are then transferred to a computer for display, review and diagnosis. A transmitted radio-frequency signal can be used to accurately estimate the location of the capsule and to track it in real time inside the body and gastrointestinal tract. It is unclear if capsule endoscopy can replace gastroscopy for those with cirrhosis.
Negative document (ID: wiki20220301en372_3997)
He co-chairs the International Conference of Capsule Endoscopy and coordinated the Consensus Conference statements to guide capsule usage. He is past president of the New York Society for Gastrointestinal Endoscopy and New York Academy of Gastroenterology and has served on the board of the American College of Gastroenterology. Lewis has written over 72 scientific papers and 27 chapters and he has authored 3 books; all in the field of gastrointestinal endoscopy. He co-authored the technical document behind the American Gastroenterological Association's position statement concerning occult and obscure gastrointestinal bleeding and the ICCE consensus statement for clinical application of capsule endoscopy. He also helped develop a scoring index for inflammatory bowel disease seen on capsule endoscopy.

Table 19: An example from MedQA-Diag dataset.

Query (ID: 225)
A 19-year-old Caucasian male presents to your office with hypopigmented skin. He undergoes a skin biopsy and is found to have an absence of melanocytes in the epidermis. What is the most likely diagnosis?
Answer
Vitiligo
Reasoning path
<p>Step 1: Identify the key clinical findings. The patient is a 19-year-old Caucasian male. He presents with hypopigmented (depigmented) areas on his skin. A skin biopsy reveals an absence of melanocytes in the epidermis.</p> <p>Step 2: Consider the differential diagnosis for hypopigmentation. Vitiligo: Autoimmune destruction of melanocytes leading to depigmented patches. Pityriasis alba: Generally shows mild hypopigmentation with retained melanocytes. Tinea versicolor: A fungal infection that causes discoloration but does not eliminate melanocytes. Nevus depigmentosus: A congenital condition with stable hypopigmentation, but melanocytes are typically present, albeit functionally abnormal.</p> <p>Step 3: Evaluate the biopsy finding. An absence of melanocytes is the hallmark of vitiligo. In other conditions causing hypopigmentation, melanocytes may be reduced or dysfunctional, but they are not completely absent.</p> <p>Step 4: Arrive at the likely diagnosis. Given the clinical presentation and the biopsy evidence (complete absence of melanocytes), the most likely diagnosis is vitiligo.</p> <p>Final Answer: Vitiligo.</p>
Positive document (ID: First_Aid_Step2_154)
<p>These spots then expand, sometimes in dermatomal patterns, to include large segments of skin. The disease is usually chronic and progressive, with some patients becoming completely depigmented. Many patients have serologic markers of autoimmune disease (e.g., antithyroid antibodies, DM, pernicious anemia) but seldom present with these diseases. Patients with malignant melanoma may develop an antimelanocyte immune response that leads to vitiligo. Diagnosed by the history and clinical picture, with histology demonstrating total absence of melanocytes. Conditions to rule out include postinflammatory hypopigmentation, scleroderma, piebaldism, and toxic exposure (pheno-lated cleansers are toxic to melanocytes).</p>
Negative document (ID: InternalMed_Harrison_4071)
<p>disorders, occurring in up to 30% of patients with vitiligo. Circulating autoantibodies are often found, and the most common ones are antithyroglobulin, antimicrosomal, and antithyroid-stimulating hormone receptor antibodies.</p>

Table 20: An example from PMC-Treatment dataset.

Query (ID: q_PMC11551046)
Case Summary: - Patient Demographics: 44-year-old female - Chief Complaint: Abdominal pain - History of Present Illness: - Presented with abdominal pain in November 2018. - In January 2019, CT scans revealed multiple pelvic masses, peritoneal and omental masses, and metastatic lesions in the liver and lungs. - Carbohydrate antigen 125 (CA125) elevated at 486.0 U/mL. - Diagnosed with stage IVb high-grade serous adenocarcinoma of the ovary in January 2019 via laparoscopic biopsy. - Past Medical History: - Hysterectomy at age 39 for uterine fibroids. - No family history of cancer. - Allergies: Not reported. - Family History: No history of cancer. - Social History: Not detailed. - Physical Exam: - Multiple metastatic lesions were documented (liver, lungs, peritoneum, and lymph nodes). - Lab Events: - Elevated CA125 at 486.0 U/mL. - Histopathological results from biopsy confirmed high-grade serous adenocarcinoma. - Imaging Events: - Initial CT scans showed extensive metastatic disease in the pelvis, peritoneum, omentum, liver, and lungs. - Subsequent imaging after chemotherapy showed significant reduction of lesions. - Final Diagnostic Results: Stage IVb high-grade serous adenocarcinoma of ovarian origin. Based on the above case summary, what would be an appropriate treatment plan for this patient?
Answer
Administer apatinib (500-750 mg orally, daily) as first-line maintenance therapy following chemotherapy and interval debulking surgery.
Reasoning path
- Objectives of the Treatment Plan: 1. Achieve tumor volume reduction through neoadjuvant chemotherapy and interval debulking surgery. 2. Prevent recurrence and extend progression-free survival through maintenance therapy. - Rationale: - The patient had advanced-stage ovarian cancer with extensive metastases requiring aggressive primary treatment using neoadjuvant chemotherapy followed by surgery. - Adjuvant chemotherapy was employed to further reduce residual disease. - Given the high costs of bevacizumab and PARPi, the patient declined these options and opted for the less expensive off-label maintenance therapy using apatinib, which is a VEGFR-2 tyrosine kinase inhibitor. This choice aligns with prior evidence supporting apatinib's anti-angiogenic efficacy in inhibiting tumor-associated vasculature.
Positive document (ID: PMC11551046_0)
Advanced ovarian cancer has a poor prognosis. In recent years, bevacizumab and PARPi have been shown to improve the prognosis of patients with advanced ovarian cancer. Tumor growth requires blood vessels to provide oxygen and nutrients. Furthermore, the vascular endothelial growth factor (VEGF) plays an important role in the process of angiogenesis. Notably, VEGF expression is higher in ovarian tumor tissues than in normal and benign ovarian tissues. There are two different clinical treatment strategies for VEGF. Bevacizumab, an anti-VEGF monoclonal antibody, inhibits the proangiogenic effect of VEGF. Another clinical treatment strategy targeting VEGF is inhibiting the function of the VEGF receptor (VEGFR). VEGF, especially VEGF-A, has been identified as a key factor in tumor angiogenesis. Bevacizumab inhibits the binding of VEGF-A to VEGFR tyrosine kinases (VEGFR1-3), inhibits tumor vascular growth, promotes tumor vascular normalization, and causes tumor cell death. Moreover, VEGFR2 is the main signaling pathway of VEGFR in vascular endothelial cells. Apatinib is a small oral VEGFR-2 tyrosine kinase inhibitor molecule that inhibits tumor angiogenesis by blocking downstream signaling.
Negative document (ID: PMC11701218_4)
Anlotinib is a multi-target tyrosine kinase inhibitor (TKI). It can inhibit various targets including vascular endothelial growth factor receptors (VEGFR), platelet-derived growth factor receptors (PDGFR), fibroblast growth factor receptors (FGFR), and c-Kit (13). It inhibits tumor cell proliferation, invasion, and metastasis and reduces tumor blood supply by inhibiting tumor angiogenesis and growth factor signaling pathways. Compared with other TKIs, anlotinib has better anti-angiogenic activity and higher selectivity, with a significantly lower incidence of grade 3 or higher adverse effects (14, 15).

Table 21: An example from PMC-Clinical dataset.

Query (ID: q_PMC3901178)
A 19-year-old girl was admitted to the gynecology clinic with the complaint of primary amenorrhea. She appears phenotypically female and has normal external genitalia and breast development, but pubic and axillary hair are absent. There are bilateral palpable masses in the inguinal regions. Trans abdominal ultrasonography reveals these masses as immature testes, and a karyotype analysis shows a 46, XY chromosomal pattern. Routine blood tests are normal, but there is slightly elevated luteinizing hormone (LH). What is the most likely diagnosis?
Answer
Complete Androgen Insensitivity Syndrome (CAIS).
Reasoning path
Let's break down the case step by step: 1. The patient is a 19-year-old with primary amenorrhea (no menstruation by age 15–16 is concerning). This sets the stage for a disorder of sexual development. 2. The patient is phenotypically female with normal external female genitalia and breast development. Normal breasts indicate that estrogen is present, which often results from the aromatization of androgens. 3. The marked absence of pubic and axillary hair is an important clue. The development of this hair is dependent on androgens. This suggests there is a problem with androgen action despite normal or elevated androgen levels. 4. Bilateral palpable masses in the inguinal regions are identified on physical exam. These are later confirmed by ultrasonography to be immature testes. Finding testes in a patient with a female phenotype is a significant finding. ... Putting all this together: The patient has a 46,XY karyotype, presence of testes, normal breast development (due to aromatization of androgens into estrogens), absence of pubic/axillary hair (due to androgen receptor insensitivity), and primary amenorrhea because she lacks a uterus. These findings are classic for Complete Androgen Insensitivity Syndrome (CAIS). Thus, the most likely diagnosis is Complete Androgen Insensitivity Syndrome (CAIS).
Positive document (ID: PMC3901178_2)
A 22-year-old woman referred to endocrinology and gynecology clinics soon after the operation on her younger sister (Case 1). Her medical history was similar to that of her sister with the symptom of primary amenorrhea. She was recently married and described no sexual problem during intercourse. She had full breast development and feminine appearance of external genitalia with sparse pubic hair. A long and blind ending vagina was found in colposcopy. There were bilateral inguinal mobile masses on palpation that resembled testes on ultrasonography. Neither uterus nor were ovaries demonstrated on the scanning of the abdomen with ultrasonography. Her karyotype was 46, XY and the level of testosterone in peripheral blood was higher than the normal female range. The other biochemical measurements were within normal limits. The patient was diagnosed as CAIS like her 19-year-old sister and her disease was explained to her with the help of a psychologist. Bilateral inguinal orchiectomy was performed in urology clinic and she was discharged at second postoperative day without complication. Histopathologic report of surgical specimen was Sertoli cell adenomas with atrophic seminiferous tubules and Leydig cell hyperplasia. ...
Negative document (ID: PMC3850340_1)
The proband was an Iranian 17 years-old girl of non-consanguineous parents with a history of irregular menstruation, every 7-8 months. No history of birth defects, mental retardation and congenital malformations was seen in the family. Pregnancy was normal and at birth the patient had no dysmorphic characteristics. In ultrasound examination, her uterine was normal but her left ovarian gonad was small and atrophic without any follicles. Right ovarian gonad was not seen. Her serum follicle-stimulating hormone (FSH) and luteinizing hormone (LH) levels were high, 151 mIU/mL and 56 mIU/mL respectively. Her height was 160 cm at the time of diagnosis. GTG banded karyotype of the patient showed a de novo reciprocal translocations between chromosomes X and 9. Her karyotype was 46, X, rcp (X; 9) (q24; q13)dn. ...

Table 22: An example from IYi-Clinical dataset.

Query (ID: 27332-1)
<p>Chief Complaint: A 56-year-old married female farmer presents with a 3-day history of cough and dyspnea, which has worsened over the past day.</p> <p>Present Illness: The patient developed dyspnea, cough, and sputum production when the weather turned cool. The sputum is minimal and difficult to expectorate. Initially, the symptoms were not severe but gradually worsened. There is no palpitation, and the symptoms are more pronounced when lying down. She has intermittently taken bronchodilator medication (details of the medication are unknown).</p> <p>Past Medical History: The patient has a 10-year history of asthma, with episodes occurring once or twice almost every year (each episode requiring several days of intravenous therapy). She is otherwise normal. She has no history of smoking or alcohol use.</p> <p>Physical Examination: Bilateral wheezing is heard on auscultation of the lungs. Heart sounds are regular at 90 beats per minute. Blood pressure is 130/80 mmHg. Respiratory rate is 20 breaths per minute.</p> <p>Ancillary Examinations: None.</p> <p>Please infer the possible initial diagnosis based on the above clinical information?</p>
Answer
Bronchial Asthma
Reasoning path
<p>Below is a systematic analysis and clinical reasoning process for this case:</p> <p>...</p> <p>Diagnostic Reasoning: The patient has a well-documented history of asthma. The current presentation, characterized by cold air-induced symptoms (cough and dyspnea) and the presence of wheezing on physical examination, strongly supports the diagnosis of bronchial asthma. The absence of other abnormal findings, along with no history of smoking or alcohol use, helps exclude alternative diagnoses such as chronic obstructive pulmonary disease (COPD), further reinforcing the likelihood of an asthma relapse.</p> <p>Conclusion: Based on the clinical history, physical findings, and prior diagnosis, a provisional diagnosis of bronchial asthma is established.</p>
Positive document (ID: 61554-1)
<p>Case Presentation: Chief Complaint: The patient was admitted to the hospital due to dyspnea and shortness of breath for over 1 month; Present Illness Over 1 month before admission, the patient developed dyspnea and shortness of breath after exposure to cold air, accompanied by coughing, which was intermittent and non-spastic, with white sticky sputum that was difficult to expectorate, chest tightness, and frequent episodes of awakening from sleep due to breathlessness. There was no chest pain, no fever, no hemoptysis, no nausea, or vomiting. Symptoms worsened after physical activity and improved with rest, being milder during the day and more severe at night. The patient received intravenous treatment at a large hospital and a local hospital, but the details were not provided.</p> <p>...</p> <p>Diagnosis Results 1. Bronchial asthma 2. Hypertension grade 2 (very high-risk group) 3. Arrhythmia Sinus tachycardia.</p> <p>Analysis and Summary: Bronchial asthma often occurs or worsens at night and/or in the early morning, and most patients can spontaneously relieve or be relieved after treatment.</p>
Negative document (ID: 66438-1)
<p>Case Presentation: Chief Complaint Recurrent cough for ten years, intermittent cough and wheezing for 2 months. Current Medical History The patient has had a cough since a cold in winter ten years ago, with little sputum, mainly dry cough. Treatment with Western medicine (specific drugs unknown) was not significantly effective. The symptoms gradually with warmer weather. Since then, the cough has recurred each winter due to colds or exposure to cold, lasting about one month, mostly dry cough, occasionally with shortness of breath, and slight relief was obtained by taking Yin-nourishing and lung-clearing pills and other drugs.</p> <p>...</p> <p>Diagnosis: Cough: Syndrome of Blood Stasis and Lung Involvement Diagnostic Basis The patient has had recurrent coughing for many years, which has damaged the lung qi and lung yin, leading to a deficiency of the body's vital energy.</p>

Table 23: The performance of retrieval models on R2MED measured by Precision@10.

Task	Size	Q&A Reference			Clinical Evidence			Clinical Case		Avg.
Model		Biology	Bioin.	MedS.	MedE.	MedD.	PMCT.	PMCC.	IiYiC.	
Sparse Retrieval										
BM25 (Robertson et al., 2009)	-	7.57	7.92	6.02	0.52	1.36	5.33	5.88	4.57	4.90
Base Size (< 1B)										
Contriever (Izacard et al., 2021)	110M	4.47	6.1	7.61	0.72	1.44	2.87	4.21	5.81	4.15
MedCPT [†] (Jin et al., 2023)	220M	0.87	7.01	3.98	0.52	0.51	1.73	3.25	3.02	2.61
InstructOR-L (Su et al., 2022)	335M	7.09	9.61	10.46	1.44	2.29	4.13	2.9	6.28	5.53
BGE-Large (Xiao et al., 2024)	335M	6.31	9.74	10	1.44	3.81	6.07	4.65	6.59	6.08
BMRetriever [†] (Xu et al., 2024)	410M	5.05	10	8.98	1.24	3.31	6.07	4.56	7.75	5.87
Large Size (> 1B)										
InstructOR-XL (Su et al., 2022)	1.5B	9.61	10.78	10.79	1.65	2.03	3.33	3.95	7.44	6.20
BMRetriever-2B [†] (Xu et al., 2024)	2B	8.06	11.3	11.59	3.3	3.9	8.2	7.02	9.15	7.82
E5-mistral (Wang et al., 2023a)	7B	8.74	14.29	12.5	2.68	5.68	6.2	8.42	10.85	8.67
BMRetriever-7B [†] (Xu et al., 2024)	7B	10.1	14.94	13.64	3.71	7.97	10.27	8.16	10.39	9.90
SFR-Embedding (Meng et al., 2024)	7B	9.13	15.07	13.98	3.92	8.39	10.2	9.91	11.16	10.22
GritLM-7B (Muennighoff et al., 2024)	7B	10.97	14.94	13.86	3.5	9.07	9.4	9.74	11.01	10.31
NV-Embed-v2 (Lee et al., 2024)	7B	11.75	16.36	14.66	3.92	8.22	10.2	10.53	6.28	10.24
Voyage-3 (Voyage AI)	-	11.26	12.47	11.82	3.3	4.41	10.07	7.54	8.84	8.71
OpenAI-3-large (OpenAI)	-	11.26	13.12	13.64	4.64	7.8	10.87	7.72	7.36	9.55

Table 24: The performance of retrieval models on R2MED measured by Recall@10.

Task	Size	Q&A Reference			Clinical Evidence			Clinical Case		Avg.
Model		Biology	Bioin.	MedS.	MedE.	MedD.	PMCT.	PMCC.	IiYiC.	
Sparse Retrieval										
BM25 (Robertson et al., 2009)	-	21.85	36.4	28.7	1.35	4.08	30.51	27.92	14.15	20.62
Base Size (< 1B)										
Contriever (Izacard et al., 2021)	110M	11.47	25.51	32.19	3.13	3.86	15.56	18.71	18.27	16.09
MedCPT [†] (Jin et al., 2023)	220M	2.81	24.74	16.84	1.7	1.29	11.1	15.86	9.26	10.45
InstructOR-L (Su et al., 2022)	335M	19.41	38.81	46.27	4.79	6.86	21.08	13.23	17.13	20.95
BGE-Large (Xiao et al., 2024)	335M	16.55	40.08	44.79	4.94	11.15	32.8	22.3	17.11	23.72
BMRetriever [†] (Xu et al., 2024)	410M	13.96	41.92	41.88	5.31	8.01	31.32	23.25	22.65	23.54
Large Size (> 1B)										
InstructOR-XL (Su et al., 2022)	1.5B	26.59	42.45	49.36	5.62	5.24	18.83	19.01	20.5	23.45
BMRetriever-2B [†] (Xu et al., 2024)	2B	23.42	43.09	49.92	12.56	11.49	42.21	34.58	28.11	30.67
E5-mistral (Wang et al., 2023a)	7B	21.57	53.49	53.12	10.4	14.11	28.68	41.3	31.49	31.77
BMRetriever-7B [†] (Xu et al., 2024)	7B	28.39	60.33	59.54	15.41	20.95	50.88	37.57	30.46	37.94
SFR-Embedding (Meng et al., 2024)	7B	22.46	55.67	58.71	15.03	21.23	50.51	48.76	31.94	38.04
GritLM-7B (Muennighoff et al., 2024)	7B	29.84	57.39	58.88	14.74	22.86	47.06	48.1	31.8	38.83
NV-Embed-v2 (Lee et al., 2024)	7B	30.95	65.35	63.53	13.65	20.18	49.32	53.07	18.32	39.30
Voyage-3 (Voyage AI)	-	29.07	48.25	55.19	14.25	12.36	52.04	38.16	27.17	34.56
OpenAI-3-large (OpenAI)	-	28.59	51.74	56.38	17.11	19.84	55.77	39.91	22.95	36.54

Table 25: Average reranking performance on R2MED using three classic rerankers: MonoBERT, BGE-Reranker, and RankLLaMA. We report nDCG@10 for three retrievers, BM25, BGE-Large, and NV-Embed-v2.

Reranker	Top-k	Biology	Bioin.	MedS.	MedE.	MedD.	PMCT.	PMCC.	IiYiC.	Avg.
BM25										
None	-	19.19	21.55	19.68	0.66	2.55	23.69	21.66	12.02	15.13
MonoBERT	10	16.12	23.57	21.45	0.93	3.21	22.61	21.25	11.50	15.08
MonoBERT	100	10.26	25.62	29.69	1.62	5.55	20.42	17.91	11.17	15.28
BGE-Reranker	10	16.61	27.26	21.87	1.18	3.67	23.79	19.29	11.21	15.61
BGE-Reranker	100	13.28	26.10	29.48	2.66	7.44	14.47	12.55	12.08	14.76
RankLLaMA	10	17.76	29.30	23.88	1.54	3.25	30.17	22.71	13.02	17.70
RankLLaMA	100	15.19	34.02	32.94	3.91	9.03	40.13	25.29	13.43	21.74
BGE-Large										
None	-	12.71	27.04	27.76	4.10	8.33	26.45	15.06	14.72	17.02
MonoBERT	10	12.73	27.15	32.25	3.33	8.16	24.03	16.38	13.50	17.19
MonoBERT	100	8.47	26.08	30.40	1.58	6.03	20.36	17.61	10.93	15.18
BGE-Reranker	10	13.86	30.91	34.81	4.17	8.61	21.45	14.68	13.49	17.75
BGE-Reranker	100	12.56	28.04	28.21	4.26	6.59	10.33	10.87	10.90	13.97
RankLLaMA	10	13.64	33.10	37.00	4.94	8.88	32.92	18.78	13.05	20.29
RankLLaMA	100	13.29	36.28	34.85	7.68	10.87	39.67	25.06	12.62	22.54
NV-Embed-v2										
None	-	27.15	50.10	47.81	10.90	16.72	44.05	39.91	14.81	31.43
MonoBERT	10	20.02	43.13	40.81	9.40	14.84	33.48	35.77	14.30	26.47
MonoBERT	100	7.43	27.01	29.66	3.03	7.90	22.84	20.49	11.21	16.20
BGE-Reranker	10	22.79	44.32	43.17	9.14	16.25	26.20	30.21	12.31	25.55
BGE-Reranker	100	14.23	28.28	28.55	5.05	8.86	6.02	12.16	9.50	14.08
RankLLaMA	10	22.55	49.38	45.07	10.56	17.24	42.66	38.36	13.64	29.93
RankLLaMA	100	15.77	36.34	30.08	8.01	13.07	39.27	27.58	12.16	22.79

Table 26: Average nDCG@10 score of generation-augmented retrieval (GAR) methods on BM25.

Method	LLM	Biology	Bioin.	MedS.	MedE.	MedD.	PMCT.	PMCC.	IYiC.	Avg.
BM25	-	19.19	21.55	19.68	0.66	2.55	23.69	21.66	12.02	15.13
HyDE	Qwen-7B	48.49	29.38	41.75	12.17	19.48	26.63	24.48	8.67	26.38
HyDE	Qwen-72B	54.42	36.96	42.12	16.05	30.84	35.95	29.30	10.89	32.07
HyDE	GPT-4o	57.34	43.02	41.22	26.55	49.60	40.61	32.80	10.42	37.70
Query2Doc	Qwen-7B	49.21	31.05	38.49	9.54	15.27	27.94	24.81	9.54	25.73
Query2Doc	Qwen-72B	57.62	39.49	41.64	15.51	28.05	34.53	31.83	10.77	32.43
Query2Doc	GPT-4o	53.32	40.34	43.41	22.03	40.74	40.26	35.57	11.16	35.85
LameR	Qwen-7B	36.70	35.18	39.67	4.60	9.40	34.92	29.51	12.55	25.32
LameR	Qwen-72B	51.03	44.79	40.75	12.98	20.51	38.78	31.62	14.74	31.90
LameR	GPT-4o	49.07	46.84	41.55	9.88	21.14	39.46	33.11	15.50	32.07

Table 27: Average nDCG@10 score of generation-augmented retrieval (GAR) methods on BGE-Large.

Method	LLM	Biology	Bioin.	MedS.	MedE.	MedD.	PMCT.	PMCC.	IYiC.	Avg.
BGE-Large	-	12.71	27.04	27.76	4.10	8.33	26.45	15.06	14.72	17.02
HyDE	Qwen-7B	19.89	30.18	40.70	12.98	15.05	41.02	17.09	15.07	24.00
HyDE	Qwen-72B	21.65	33.96	41.16	16.01	19.82	44.03	21.67	16.80	26.89
HyDE	GPT-4o	22.59	32.97	41.29	19.45	27.18	45.43	23.28	16.85	28.63
Query2Doc	Qwen-7B	27.70	34.56	42.80	11.65	14.84	23.79	17.49	14.73	23.45
Query2Doc	Qwen-72B	34.86	41.70	47.44	15.86	28.61	25.85	24.39	14.79	29.19
Query2Doc	GPT-4o	33.23	40.20	44.92	19.14	37.96	29.16	26.31	14.85	30.72
LameR	Qwen-7B	20.33	38.37	41.70	7.17	14.86	44.64	20.22	16.21	25.44
LameR	Qwen-72B	22.60	37.73	43.99	15.32	20.38	45.57	32.75	20.59	29.87
LameR	GPT-4o	27.14	42.24	44.86	12.75	21.65	42.94	32.97	21.40	30.74

Table 28: Average nDCG@10 score of generation-augmented retrieval (GAR) methods on NV-Embed-v2.

Method	LLM	Biology	Bioin.	MedS.	MedE.	MedD.	PMCT.	PMCC.	IYiC.	Avg.
NV-Embed-v2	-	27.15	50.10	47.81	10.90	16.72	44.05	39.91	14.81	31.43
HyDE	Qwen-7B	30.12	49.95	49.39	13.37	19.49	42.99	38.36	17.86	32.69
HyDE	Qwen-72B	32.35	54.43	49.97	18.83	24.47	47.40	44.88	23.68	37.00
HyDE	GPT-4o	33.61	54.15	50.83	23.08	36.09	47.35	48.51	21.30	39.37
Query2Doc	Qwen-7B	33.04	50.62	48.34	15.76	17.44	40.00	33.51	21.41	32.52
Query2Doc	Qwen-72B	37.09	52.13	51.38	21.08	30.01	42.78	43.81	24.03	37.79
Query2Doc	GPT-4o	37.01	52.35	51.95	25.86	43.30	47.10	52.30	23.37	41.66
LameR	Qwen-7B	34.26	50.19	51.73	9.38	17.64	48.96	44.70	19.27	34.52
LameR	Qwen-72B	36.52	55.48	52.94	20.00	28.64	50.20	44.48	20.25	38.56
LameR	GPT-4o	39.32	57.25	53.84	14.53	27.05	48.86	44.70	19.23	38.10

Table 29: The nDCG@10 performance of large reasoning models on BM25.

Model	Biology	Bioin.	MedS.	MedE.	MedD.	PMCT.	PMCC.	IYiC.	Avg.
BM25	19.19	21.55	19.68	0.66	2.55	23.69	21.66	12.02	15.13
Large Language Model									
Qwen2.5-7B-Ins.	48.49	29.38	41.75	12.17	19.48	26.63	24.48	8.67	26.38
Qwen2.5-32B-Ins.	52.88	39.2	42.42	16.8	25.87	33.27	26.72	11.81	31.12
Llama3.1-70B-Ins.	52.54	39.42	41.05	16.99	33.87	37.32	28.67	9.32	32.40
GPT-4o	57.34	43.02	41.22	26.55	49.6	40.61	32.8	10.42	37.70
Large Reasoning Model									
R1-Distill-Qwen-32B	48.9	38.8	38.28	16.04	25.21	31.55	22.47	11.15	29.05
QwQ-32B	58.24	42.35	42.2	23.7	38.12	41.65	26.66	7.34	35.03
R1-Distill-Llama-70B	49.52	38.86	39.48	22.12	38.4	34.95	33.86	9.13	33.29
HuatuoGPT-o1-70B	49.77	40.02	38.47	21.91	39.54	39.3	28.38	10.01	33.43
o3-mini	59.65	46.56	47.17	34.64	55.22	41.65	35.32	7.86	41.01
Search-Enhanced Large Reasoning Model									
Search-R1-3B	34.38	20.31	30.04	5.47	13.18	31.66	19.69	5.53	20.03
Search-R1-7B	36.51	26.92	34.62	10.02	13.88	34.98	28.18	11.37	24.56
Search-o1_QwQ-32B	60.78	41.58	43.93	24.07	39.94	37.52	28.38	8.34	35.57
Search-o1_Qwen3-32B	56.84	39.48	43.01	22	34.42	41.62	26.2	7.16	33.84

Table 30: The nDCG@10 performance of large reasoning models on BGE-Large.

Model	Biology	Bioin.	MedS.	MedE.	MedD.	PMCT.	PMCC.	IYiC.	Avg.
BGE-Large	12.71	27.04	27.76	4.1	8.33	26.45	15.06	14.72	17.02
Large Language Model									
Qwen2.5-7B-Ins.	19.89	30.18	40.7	12.98	15.05	41.02	17.09	15.07	24.00
Qwen2.5-32B-Ins.	23.6	33.76	41.51	14.41	17.98	39.77	22.23	16.22	26.19
Llama3.1-70B-Ins.	21.4	37.06	41.44	15.91	22.85	45.78	24.61	16.36	28.18
GPT-4o	22.59	32.97	41.29	19.45	27.18	45.43	23.28	16.85	28.63
Large Reasoning Model									
R1-Distill-Qwen-32B	19.61	33.42	40.76	14.78	19.3	42.38	23.84	17.1	26.40
QwQ-32B	21.01	32.89	39.45	16.68	22.55	43.4	20.75	15.95	26.59
R1-Distill-Llama-70B	21.68	34.79	40.33	19.06	23.93	42.74	26.44	18.28	28.41
HuatuoGPT-o1-70B	18.88	35.5	39.72	22	27.48	43.29	26.46	18.5	28.98
o3-mini	22.18	37.62	43.97	25.39	34.91	44.55	24.65	17.04	31.29
Search-Enhanced Large Reasoning Model									
Search-R1-3B	16.8	26.83	36.74	8.4	12.17	37.15	22.84	16.16	22.14
Search-R1-7B	19.14	30.38	37.31	12.53	14.81	38.93	24.51	20.6	24.78
Search-o1_QwQ-32B	21.92	32.34	42.71	20.1	24.34	42.91	26.31	18.66	28.66
Search-o1_Qwen3-32B	23.88	30.34	42.16	18.48	25.49	42.96	25.46	17.47	28.28

Table 31: The nDCG@10 performance of large reasoning models on OpenAI-3-large.

Model	Biology	Bioin.	MedS.	MedE.	MedD.	PMCT.	PMCC.	IYiC.	Avg.
OpenAI-3-large	23.82	40.51	44.05	11.78	15.01	47.43	28.87	17.12	28.57
Large Language Model									
Qwen2.5-7B-Ins.	30.15	42.33	45.79	15.45	19.73	48.64	30.39	19.79	31.53
Qwen2.5-32B-Ins.	31.37	45.46	46.44	21.13	24.69	48.05	34.45	22.38	34.25
Llama3.1-70B-Ins.	31.31	46.83	48.14	21.42	28.32	51.22	37.11	21.7	35.76
GPT-4o	32.15	45.99	47.97	27.28	36.92	51.24	38.96	21.82	37.79
Large Reasoning Model									
R1-Distill-Qwen-32B	30.39	44.89	47.62	20.33	27.89	49.04	36.26	21.04	34.68
QwQ-32B	32.51	45.46	46.76	24.81	31.76	52.78	35.1	22.89	36.51
R1-Distill-Llama-70B	30.08	47.44	48.98	24.94	33.24	50.49	41.39	22.32	37.36
HuatuoGPT-o1-70B	29.77	47.7	48.4	27.46	36.79	52.93	39.26	23.63	38.24
o3-mini	31.5	48.61	48.19	31.74	39.51	51.69	38.92	22.56	39.09
Search-Enhanced Large Reasoning Model									
Search-R1-3B	24.48	40.91	47.65	11.62	16.78	47.79	32.27	20	30.19
Search-R1-7B	28.3	43.6	48.03	15.2	18.57	48.38	38.15	22.92	32.89
Search-o1_QwQ-32B	32.43	46.95	47.84	25.19	31.22	51.67	36.58	26.51	37.30
Search-o1_Qwen3-32B	34.73	45.38	47.18	26.95	33.71	50.78	40.02	24.17	37.87