DEEPMED: Building a Medical DeepResearch Agent via Multi-hop Med-Search Data and Turn-Controlled Agentic Training & Inference

Zihan Wang[♠,♡*] **Hao Wang**[♡*] **Shi Feng**[♠†] **Xiaocui Yang**[♠] **Daling Wang**[♠]
Yiqun Zhang[♠] **Jinghao Lin**[♡] **Xiaozhong Ji**[♡] **Haihua Yang**[♡]
[♠]School of Computer Science and Engineering, Northeastern University, Shenyang 110819, China,
[♡]ByteDance
{2310744@stu.neu.edu.cn, fengshi@cse.neu.edu.cn}

Abstract

Medical reasoning models remain constrained by parametric knowledge and are thus susceptible to forgetting and hallucinations. DeepResearch (DR) models ground outputs in verifiable evidence from tools and perform strongly in general domains, but their direct transfer to medical field yields relatively limited gains. We attribute this to two gaps: task characteristic and tool-use scaling. Medical questions require evidence interpretation in a knowledge-intensive clinical context; while general DR models can retrieve information, they often lack clinical-context reasoning and thus "find it but fail to use it," leaving performance limited by medical abilities. Moreover, in medical scenarios, blindly scaling tool-call can inject noisy context, derailing sensitive medical reasoning and prompting repetitive evidence-seeking along incorrect paths. Therefore, we propose DEEPMED. For data, we deploy a multi-hop med-search QA synthesis method supporting the model to apply the DR paradigm in medical contexts. For training, we introduce a difficulty-aware turn-penalty to suppress excessive tool-call growth. For inference, we bring a monitor to help validate hypotheses within a controlled number of steps and avoid context rot. Overall, on seven medical benchmarks, DEEPMED improves its base model by 9.79% on average and outperforms larger medical reasoning and DR models.

1 Introduction

In the medical field, reasoning models (Liu et al., 2025b; Zhang et al., 2025b; Huang et al., 2025) can achieve stronger diagnostic performance than conventional LLMs, while they remain constrained by the limits of parametric knowledge (Chen et al., 2024). It can induce hallucinations and spurious attributions, leading to diagnostic bias and potential

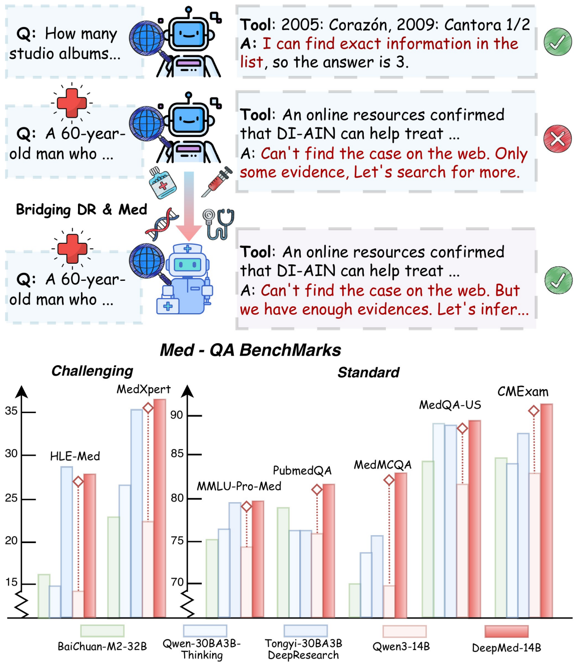

Figure 1: General search tasks typically provide definitive answers online, but medical tasks are different; while online information can assist, the final answer relies on medical reasoning **(up)**. DEEPMED shows significant improvements over its base reasoning model on medical benchmarks, and outperforms larger medical reasoning model and general DR models **(down)**.

clinical risks (Feng et al., 2025; Thirunavukarasu et al., 2023). While the emerging DeepResearch (DR) paradigm in general domains tightly integrates web search with model reasoning, grounding generation in verifiable external evidence and thereby reducing hallucinations (Li et al., 2025b,c). DR has shown strong capability in complex problem solving and multi-step reasoning, particularly in open-domain question answering (QA) (Wei et al., 2025; Zhou et al., 2025). Nevertheless, directly transferring general-purpose DR models to medical settings remains highly challenging.

We evaluated Tongyi-DeepResearch (Team et al., 2025b), a strong DR model trained with large-scale agentic training, on medical benchmarks and found

*Equal contribution.
†Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 18160–18178
July 2-7, 2026 ©2026 Association for Computational Linguistics

that its gains in medical tasks are modest against its base reasoning model - Qwen3-30BA3B-Thinking as shown in Fig. 1, especially when contrasted with its substantial improvements on general search tasks. Two key gaps are identified in translating DR capabilities to medical diagnosis. **(i) Task characteristic mismatch.** General DR models training primarily targets general factual queries (e.g., dates, people, and entities) (Wei et al., 2025; Zhou et al., 2025). In contrast, medical questions require differential reasoning and decision making in a knowledge-intensive clinical context: the model must not only retrieve information, but also interpret its clinical implications, weigh the strength and conflicts of evidence, and use medical priors to filter key cues. **(ii) Tool-scaling profit mismatch.** During both training and inference, general DR systems often improve performance by scaling the number of tool calls to gather more information (Li et al., 2025b,c) like Fig. 1. In medical diagnosis, however, more is not necessarily better. Excessive searching introduces noise that accelerates context rot (Hong et al., 2025), degrading the context—an effect especially pronounced in medicine due to the sensitivity and error-proneness of medical information (Neha et al., 2025; Ke et al., 2025), which ultimately constrains model performance. Furthermore, injecting large medical context can trigger an "over-evidence" phenomenon: where the model struggles to integrate the accumulated information and thus repeatedly calls tools to re-verify what it already knows instead of advancing its reasoning, creating a self-reinforcing negative feedback loop.

To migrate these gaps, we propose DEEPMED, a DeepResearch model tailored for medical scenarios. We introduce a web-based method to synthesize multi-hop medical questions used for agentic Supervised Fine-Tuning (SFT) stage, it strengthen multi-hop search ability in the medical knowledge-rich context. During agentic Reinforcement Learning (RL), we train the model using challenging diagnostic tasks to further improve its ability to integrate search and medical reasoning capabilities. To control the unrestrained tool scaling often found in Agentic RL (Feng et al., 2025), we deploy a difficulty-aware turn-penalty to implicitly suppress it while training. For easier samples, the model is encouraged to make fewer tool calls to reduce redundant exploration; for harder samples, it is allowed a moderate increase in tool usage to acquire crucial evidence. For inference, we deploy a monitor to explicitly rescue the model from "over-evidence". When the model begins to repeatedly validate a candidate answer, it halt further tool exploration in time and help it to finalize the response. Ultimately, DEEPMED achieves substantial average improvements over the base model, Qwen3-14B, as shown in Fig. 1.

Our contributions can be summarized as follows:

- We identify two gaps that arise when applying DR paradigm to medical field—one in task characteristics and the other in the benefits of tool scaling—and address them with DEEPMED.
- DEEPMED improves by an average of **13.92%** on the two challenging benchmarks and **8.13%** on the five standard benchmarks against its base model and outperforms several medical reasoning models and DR models with more parameters or trained on more data.
- We validate that DEEPMED can leverage web medical evidence to reduce hallucinations and self-correct in reasoning, demonstrating the promise of DeepResearch in the medical field.

2 Related Work

2.1 DeepResearch Models

DeepResearch models utilize web search tools and real-world network knowledge to find or verify the information needed for inference (Zhang et al., 2025a). OpenAI DeepResearch is the most well-known example, and this capability has become a target for improvement for various research institutions and companies (Zeng et al., 2025; Team et al., 2025a). Early works such as Webwalker/Search-R1 (Wu et al., 2025a; Jin et al., 2025) explore the capabilities of this paradigm, achieving significant improvements on some search benches. Subsequent works such as Websailor/Deepdive (Lu et al., 2025b; Li et al., 2025c) further validate that multi-hop QA problems and agentic RL are key to unlocking these capabilities. This paradigm exhibits strong generalization, bringing improvements even to extremely hard QA problems such as Humanity-Last-Exam (HLE) (Phan et al., 2025a). MedResearcher-R1 (Yu et al., 2025a) first applies this capability to MedBrowseComp (Chen et al., 2025), a medical search benchmark that does not target diagnostic or clinical decision-making performance. Accordingly, we stress that the central challenge is bridging search and medicine: medical systems must go beyond better retrieval to use retrieved evidence to drive and verify diagnostic hypotheses.

2.2 Medical Reasoning Models

LLMs are widely used in medicine (Liu et al., 2024; Wu et al., 2024; Qiu et al., 2024; Liu et al., 2025c), and stronger chain-of-thought reasoning often brings substantial performance gains. HuatuoGPT-o1 (Chen et al., 2024) trains a 70B model with SFT followed by RL and achieves better performance on medical reasoning benchmarks. In parallel, M1 (Huang et al., 2025), AlphaMed (Liu et al., 2025b), and BioMed-R1 (Thapa et al., 2025) focus on pushing the upper bound of medical reasoning via reinforcement learning. Industrial efforts such as Baichuan-M2 (Dou et al., 2025) and Quark-Med (Li et al., 2025a) further demonstrate the promise of medical reasoning models. Despite these advances, most methods still rely on internal reasoning, making them prone to forgetting and hallucinations. Combining medical reasoning with external knowledge and tool use is therefore key for robustness and real-world utility. While RAG-style methods help (Xiong et al., 2024; Zhao et al., 2025), their dependence on private databases and loose coupling between reasoning and dynamically retrieved evidence limit multi-hop solving (Jin et al., 2025). In contrast, the open web offers abundant reliable medical information, highlighting the promise of DR models for medicine.

3 Preliminaries

DeepResearch Paradigm Standard "thinking" models implement slow thinking by first producing a chain of thought (CoT) (Jaech et al., 2024; Guo et al., 2025) wrapped in <think> tags to analyze the input problem x, and then output the final answer y. The resulting trace can be written as $\mathcal{T} = \langle x, (\mathrm{CoT}), y \rangle$. In contrast, DeepResearch models interleave tool use with the reasoning process, typically following the ReAct (Yao et al., 2022) paradigm. After <think>, tool invocations are wrapped in <tool_call> tags. A downstream tool server executes the call and returns the result wrapped in <tool_response>, which is then appended to the model context. This yields a multi-turn interaction, and the model outputs the final answer only after it stops calling tools.

$$\begin{aligned}\mathcal{T} = \langle x, &\underbrace{(\mathrm{CoT}_1, \mathrm{Tool}_1^{\mathrm{call}}, \mathrm{Tool}_1^{\mathrm{res}})}_{\text{Turn } 1}, \dots, \\ &\underbrace{(\mathrm{CoT}_N, \mathrm{Tool}_N^{\mathrm{call}}, \mathrm{Tool}_N^{\mathrm{res}})}_{\text{Turn } N}, y \rangle,\end{aligned} \tag{1}$$

Tool Configuration We leverage web-based external knowledge—preferred for its broad coverage and accessibility—as the primary grounding source for DEEPMED. Specifically, the model accesses online information through two tools, `Search` and `Visit`. `Search` takes a `query` as input and returns a list of relevant URLs along with brief snippets. `Visit` takes a target `Url` and a `Goal` as input, retrieves the full webpage content, and then returns a condensed summary of the information that is most relevant to the specified goal (Team et al., 2025b).

4 DEEPMED Methodology

In this section, we formalize the key components of DEEPMED and present the underlying methodology with equations. Our training procedure consists of two stages: (i) Warm-up Agentic SFT (ASFT) with Multi-hop Med-Search QA, and (ii) Agentic RL (ARL) with Hard Med-Diagnosis QA. The first stage teaches the model to invoke tools properly in a medically rich context, while the second stage further couples retrieval with medical reasoning, unlocking the model's potential on medical tasks. For inference, we introduce (iii) Inference with Over-Evidence Monitor, which prevents the model from falling into the over-evidence trap.

4.1 ASFT with Multi-hop Med-Search QA

Agentic SFT is a critical stage in training DeepResearch models. It serves to warm-up the model, reshape its reasoning paradigm, and teach it to interface effectively with external knowledge sources by tools. Prior work has consistently shown that skipping this stage typically results in degraded performance (Team et al., 2025b; Wu et al., 2025a).

4.1.1 Multi-hop Med-Search QA Synthesis

Multi-hop Medical Chain Construction The essence of multi-hop QA lies in constructing an entity chain that encourages the model to decompose a question step by step along a logical path before arriving at the final answer. In this work, we synthesize such medical chains using web-based knowledge as shown in Fig. 2. Compared with knowledge-graph-based approaches (Lu et al., 2025b; Li et al., 2025c), our web-based method is more flexible and scalable, and it can leverage diverse online sources to build more reliable medical links. Concretely, we start from a medical keyword or entity (e.g., a disease or a drug). We collect information about the entity from authoritative websites, extract relevant medical facts, and consolidate them

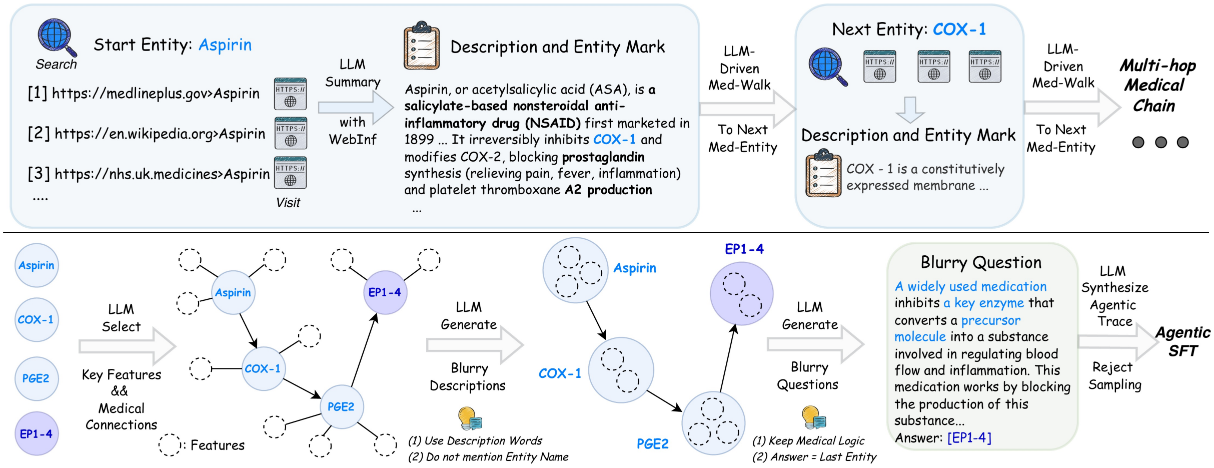

Figure 2: The multi-hop Med-Search data synthesis workflow for Agentic SFT of DEEPMED. The top panel illustrates how we synthesize multi-hop medical chains that emphasize logical relations by leveraging multi-source web evidence. The bottom panel shows how these chains are transformed into Med-Search QA.

into a short entity summary. Then identify candidate related nodes from the summary and, guide by medical logic (e.g., key disease characteristics, indicated treatments, or a drug's key ingredients and targets), select the most relevant entity as the next hop. Repeating this process forms a multi-hop sequence, where each hop is connected by verifiable medical relations. By iteratively traversing these links, we can flexibly extend the number of hops and ultimately construct a multi-hop medical chain from a start entity to an end entity.

Med-Search QA Generation After constructing the medical chain, we translate its underlying logical relations into a multi-hop QA problem that requires step-by-step reasoning. A central design principle is entity obfuscation (identity blurring): we prohibit explicitly mentioning any major entity names in the question. As a result, the model must rely on tool use and medical reasoning to infer entity identities from functional roles and interactions, recover the latent chain, and ultimately derive the answer. To achieve this, we use an LLM (e.g., Gemini-2.5-pro) to obfuscate each entity description by paraphrasing it into an anonymized form that preserves only key distinguishing attributes. We then prompt the LLM to compose these obfuscated descriptions into a coherent Med-Search QA that faithfully reflects the original logical chain.

Quality and Difficulty Control For quality control, we use strong models (e.g., GPT-5) to filter the synthesized questions using full contextual information, focusing primarily on validating the soundness of the underlying logical chains. To increase difficulty, we deliberately extend the chain length (i.e., the number of hops), making the reasoning more complex and involving more entities. After question synthesis, we apply a pass@4 filter using Qwen-14B with a single tool call. We retain only questions that the model answers correctly at most once (i.e., 0/4 or 1/4), ensuring that the selected questions cannot be reliably solved by internal reasoning alone or by a trivial single-step retrieval. A case question is shown in Appendix. A.1.

4.1.2 Warm-up ASFT

Tools are wrapped as an MCP server and paired with an existing model to synthesize solution trajectories. Rejection sampling is then applied to retain only correct trajectories with a reasonable number of tool-use rounds. The resulting trajectories follow the format in Eq. 1. For training, a multi-turn dialogue-style formulation is used: the loss on non-model-generated segments (i.e., Tool^{res}) is masked out, and the standard SFT objective is applied to all remaining tokens, as defined in Eq. 2.

$$\mathcal{L}_{\text{SFT}}(\theta) = -\sum_{t=1}^{|\mathcal{T}|} M_t \cdot \log \pi_\theta(a_t \mid a_{<t}, x)\,, \quad (2)$$

a denotes the tokens of the trajectory and $M_t \in \{0, 1\}$ is a binary mask. Specifically, $M_1 = 0$ if token a_t falls within x or Tool^{res}, and $M_t = 1$ for tokens within CoT and $\text{Tool}^{\text{Call}}$.

4.2 ARL with Hard Medical QA

ASFT teaches the model to use tools effectively in medical settings, delivering substantial gains over the base model. Building on this, we apply Agentic

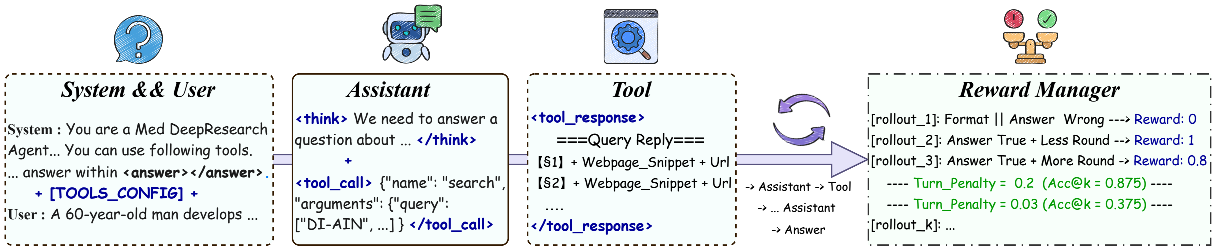

Figure 3: A full rollout and its reward evaluation are shown. Only segments inside the solid boxes contribute to the loss. Incorrect format or answer receives zero reward, and even correct rollouts are penalized for excessive rounds.

RL (ARL) on diagnosis-oriented problems to further strengthen medical search–reasoning integration and performance. Fig. 3 shows the workflow.

4.2.1 Challenging Medical QA Filter

Following prior work (Liu et al., 2025b; Huang et al., 2025), we curate a hard subset from the training split of existing datasets (e.g., MedQA-USMLE). The filtering proceeds in two stages: (i) we run Qwen-14B with direct (tool-free) reasoning for four trials and discard any questions with accuracy strictly above 50%. For remaining questions, (ii) we run Qwen-14B with a simple tool-assisted setup (typically a single tool call) for four trials, and keep only those with accuracy 0 or 0.25.

4.2.2 ARL with Difficult-aware Turn-Penalty

We utilize GRPO (Shao et al., 2024) as the RL algorithm. As in ASFT, only the model-generated components (i.e., CoT and $\text{Tool}^{\text{call}}$) participate in optimization in ARL. For each medical query q, we sample a group of G outputs from the old policy. Following prior work (Lu et al., 2025b; Yu et al., 2025b), we remove the KL-divergence regularization term. The policy is updated by maximizing the objective as Equ. 3.

$$\mathcal{L}_{\text{GRPO}}(\theta) = \frac{1}{G}\sum_{i=1}^{G} \min\left(\rho_i \hat{A}_i, \text{clip}(\cdot)\hat{A}_i\right) \quad (3)$$

where ρ_i is the importance sampling ratio, and *clip* is the clip range. The advantage $\hat{A}_i$ is computed by normalizing the rewards within the group average:

$$\hat{A}_i = \frac{r_i - \text{mean}(r)}{std(r)} \quad (4)$$

where r_i is the total reward for trajectory y_i.

Reward Design In standard GRPO, rewards depend only on answer correctness, whereas in agentic training, format error (e.g., invalid tool calls, over long) are also treated as incorrect (Team et al., 2025b). Meanwhile, we observe an "over-evidence" behavior, where the model keeps invoking tools redundantly despite already being able to answer, which yields low-quality rollouts and hinders policy learning. To address this, we introduce a turn-penalty for successful but overly long rollouts. Our overall reward is defined as Equ. 5

$$r_i = \begin{cases} 1, & F \wedge A \wedge T \\ 1 - \lambda\, r_i^{\text{turn}}, & F \wedge A \wedge \neg T \\ 0, & \neg F \ \text{ or } \ \neg A \end{cases} \quad (5)$$

where (F, A) indicates whether the trajectory passes the format and answer checks, and T is the turn-budget check defined as $T \triangleq \left(N_i \leq \bar{N}^G_{suc}\right)$. N_i is the number of interaction rounds and $\bar{N}^G_{\text{suc}}$ is the mean turn count over true rollouts in the group G. λ controls the penalty strength. The turn-penalty is defined as:

$$r_i^{\text{turn}} = \omega f\left(N_i - \bar{N}^G_{\text{suc}}\right) \quad (6)$$

Here, $f(\cdot)$ denotes a smoothing function; in our implementation, we use the logarithm. The coefficient ω is difficulty-adaptive: we estimate problem difficulty using the model's average rollout accuracy and assign smaller ω to harder questions.

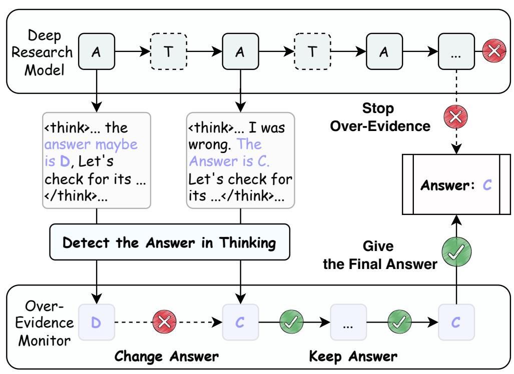

Figure 4: The workflow of 'Over-Evidence' monitor.

4.3 Inference with Over-Evidence Monitor

Although med-enhanced agentic RL has substantially reduced the prevalence of "Over-Evidence",

Models/Datasets	HLE-Med	MedXpert	MMLU-Pro-Med	PubMedQA	MedMCQA	MedQA-USMLE	CMExam
Awesome General Models							
Gemini2.5-Pro	14.36	39.27	84.30	76.00	84.13	92.22	87.40
Kimi-K2-Thinking-1TB (Team et al., 2025a)	18.79	41.59	82.74	76.60	80.30	94.34	88.47
Deepseek-v3.2-685B (Liu et al., 2025a)	19.46	43.84	83.19	75.80	81.38	93.72	91.84
Qwen3-30BA3B-Thinking (Yang et al., 2025)	14.77	26.40	76.94	76.00	73.11	88.14	83.81
Qwen3-14B (Yang et al., 2025)	12.75	22.40	74.53	75.40	68.18	82.17	82.79
Medical Reasoning Models							
HuatuoGPT-o1-70B (Chen et al., 2024)	11.60	26.50	76.61	80.60	73.61	83.27	70.42
M1-1K-7B (Huang et al., 2025)	14.77	16.70	65.73	77.50	58.26	71.56	70.64
M1-1K-32B (Huang et al., 2025)	16.78	25.20	71.53	77.60	67.37	83.50	79.03
MedResaon-8B (Wu et al., 2025b)	22.40	16.40	63.13	77.60	57.25	68.40	68.95
AlphaMed-70B (Liu et al., 2025b)	-	32.56	79.54	80.90	75.89	87.52	-
BaiChuan-M2-32B (Dou et al., 2025)	16.78	22.80	75.05	79.20	68.49	84.68	78.73
QuarkMed-32B (Li et al., 2025a)	-	28.60	-	79.00	75.50	86.02	88.61
DeepResearch Models							
Tongyi-DeepResearch (Team et al., 2025b)-30BA3B	**28.19**	35.40	79.54	76.20	75.21	87.90	87.59
DEEPMED-14B-SFT (ours)	19.46	34.80	77.92	80.40	79.37	87.27	90.31
DEEPMED-14B-RL (ours)	26.84	**36.14**	**79.93**	**81.80**	**82.60**	**88.22**	**91.19**

Table 1: Performance comparison of DEEPMED and baselines on medical benchmarks ("-" indicates not reported).

the model remains prone to this failure mode when confronted with particularly challenging problems. To maintain stable inference, we introduce a lightweight sub-agent that monitors OE during inference, as shown in Fig. 7. At each step, it reads the DR model's response and infers its intent. Then it caches any proposed answer; if the cached answer remains unchanged in the next turn, the model is likely invoking tools solely for detail verification. Limited validation can be beneficial, but repeated validation without any cache updates signals stalled progress and merely injects noise into the context. Accordingly, if the cache remains unchanged for more than a preset number of turns, the monitor terminates the interaction and returns the cached answer as the model's final output. This early-stopping fallback is preferable to letting the model continue iterating under an increasingly corrupted context—where redundant tool calls bloat the prompt, exacerbate context rot, and may end only when the turn budget is exhausted.

5 Experiments

5.1 Settings

Baselines and Datasets. We select a total of 13 models for our baselines spanning three groups: Awesome General Models, Medical reasoning models, and DeepResearch models. We compare on seven plain-text datasets, including medical subsets of several general benchmarks. Details of them are provided in the Appendix. A.2.

Training Settings. Our model is trained based on the Qwen3-14B model. We use 5,437 multi-hop med-search data for ASFT and 6,304 difficulty-filtered medical data for ARL. More settings are located in Appendix. A.3.

Tools Configuration. We equip the model with a `search` and a `visit` tool. The former uses query-based search, returning multiple webpage URLs and brief descriptions, while the latter is responsible for retrieving the necessary information from the full content of a specific webpage. Tool details are in Appendix. A.3.

5.2 Main Results

Tab. 1 presents the performance comparison between DEEPMED and other models on seven medical benchmarks. We can summarize three insights: **(ii)** After Warm-up SFT, DEEPMED's performance on many benchmarks has surpassed most medical expert models with more parameters, especially the more difficult HLE-Med and MedXpert. On some common medical benchmarks, DEEPMED even surpassed Tongyi-DR. This demonstrates that pure medical multi-hop problems significantly enhance the model's ability in medical diagnosis. **(ii)** After the RL stage, performance improves across all benchmarks. DeepMed achieves state-of-the-art results on every benchmark except HLE-Med. Notably, it does so with fewer parameters and substantially less medical training data, underscoring the strong potential of deepresearch capabilities for medical diagnosis.. **(iii)** General-purpose Tongyi-DR and strong medical LLMs like AlphaMed and QuarkMed each outperform the other on different benchmarks, suggesting that search capability and medical reasoning offer complementary strengths. DEEPMED is the bridge of them.

5.3 Ablation Study

In ablation experiments, we mainly need to address the following three questions: (i) the qual-

Model	HLE-Med		MedXpert		MMLU-Pro		PubMedQA		MedMCQA		MedQA		CMExam	
	#Acc	#Turn	#Acc	#Turn	#Acc	#Turn	#Acc	#Turn	#Acc	#Turn	#Acc	#Turn	#Acc	#Turn
DeepMed-SFT-14B	19.46	10.4	34.80	8.8	77.92	5.8	80.40	2.7	79.37	4.8	87.27	4.9	90.31	5.3
DeepMed-SFT-14B w GENERAL	18.12	12.2	32.89	12.0	76.71	7.2	78.80	4.0	76.88	5.4	86.72	5.8	88.69	6.5
DeepMed-RL-14B w/o Penalty	26.84	14.6	35.52	14.7	78.63	9.2	79.80	5.4	79.44	6.7	87.88	6.7	90.89	7.3
DeepMed-RL-14B w/o monitor	24.83	12.9	35.72	11.2	77.61	8.2	81.20	3.4	80.11	7.8	86.88	6.4	90.01	6.4
DeepMed-RL-14B	**26.84**	11.8	**36.14**	9.4	**79.93**	6.2	**81.80**	2.8	**82.60**	5.2	**88.22**	5.0	**91.19**	4.7

Table 2: Performance comparision for ablation results: w/o Monitor removes the OE Monitor; w/ General uses general training data; w/o Penalty removes the turn-penalty.

ity of our synthesized multi-hop med-search data; (ii) the benefits of difficult-aware turn-penalty in ARL; and (iii) whether the OE Monitor improves performance. So we have three variants: (i) a model using open-source general multi-hop data from Deepdive (Lu et al., 2025b) for SFT; (ii) an RL model without turn-penalty; and (iii) removing the OE Monitor from the RL model. Results in Tab. 2 answers the question above: removing any component degrades DeepMed's performance, indicating that each of our proposed improvements contributes meaningfully to final results.

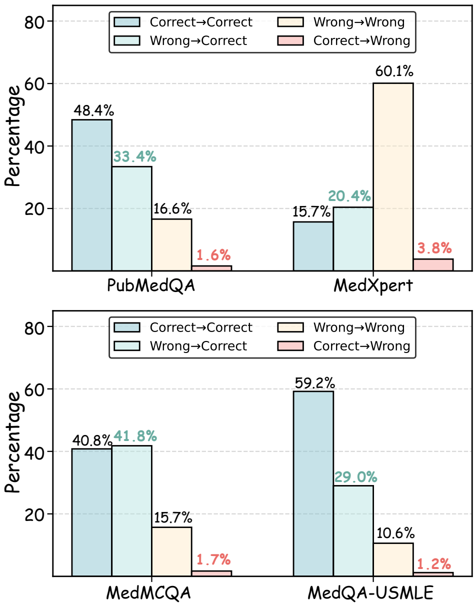

Figure 5: Answer outcome breakdown: most correct predictions stem from tool use and self-correction, while errors caused by later tool use or reasoning are rare.

5.4 Medical Evidence-Seeking Ability Analyse

In our experiments, we find that under the deep-research paradigm, DeepMed typically first proposes a tentative answer and then uses search to verify its details, revise the hypothesis, and repeat this search–reflection loop until it converges on a final decision. To quantify the effectiveness of this search-and-reflection process, we analyze how the model's initial answer differs from its final answer. Specifically, we categorize each instance into four cases: (i) initially correct and finally correct; (ii) initially incorrect but corrected during search, ending correct; (iii) incorrect throughout; and (iv) initially correct but revised to an incorrect final answer.

As shown in the Fig. 5, across all four evaluation groups, a large fraction of initially incorrect answers are corrected to become correct, while cases that flip from initially correct to finally incorrect do occur but are much rarer. These results suggest that DeepMed can effectively exhibit an evidence-based pattern of medical reasoning, using retrieved evidence to revise erroneous hypotheses. An illustrative case is in Appendix A.5.

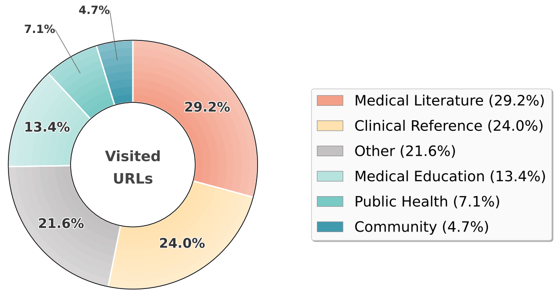

Figure 6: Distribution of website categories visited by DeepMed via the `Visit` tool.

Category	Representative Root Domains
Medical Literature	`pubmed.ncbi.nlm.nih.gov` `nejm.org`
Clinical Reference	`uptodate.com` `mayoclinic.org`
Medical Education	`amboss.com` `medbullets.com`
Public Health	`cdc.gov` `nih.gov`
Medical Community	`researchgate.net` `forums.studentdoctor.net`

Table 3: Representative root domains for each category of visited web sources.

5.5 Information Source Details

Fig. 6 shows the distribution of website categories accessed via `Visit` during training, and Tab. 3 lists representative root domains. Most visits come from authoritative sources, dominated by Medical Literature (29.2%; e.g., PubMed/PMC/NEJM) and Clinical References (24.0%; e.g., UpToDate/-Mayo Clinic/Merck Manuals). Medical Education sites account for 13.4% (e.g., AMBOSS, MedBullets), while Public Health (7.1%) and Medical Community platforms (4.7%) play supplementary roles. Notably, even within the *Other* category, a large fraction of visits still come from high-quality knowledge bases such as Wikipedia. Overall, the distribution indicates that DEEPMED primarily relies on expert-curated evidence, supporting the reliability of retrieved web knowledge.

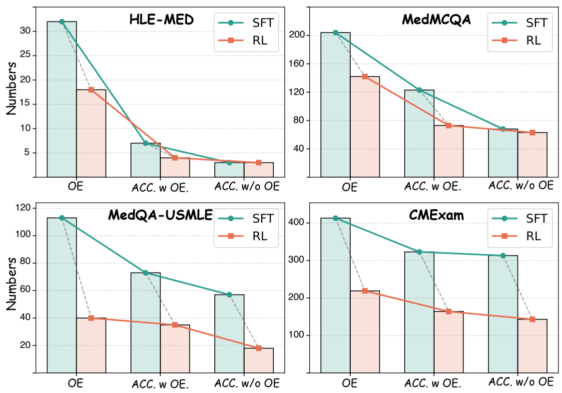

Figure 7: An illustration of DEEPMED-SFT/RL on four datasets: OE Monitor (OEM) trigger counts and the number of correct predictions with OEM versus correct predictions continued inference without OEM.

5.6 Over-Evidence Monitor Analysis

Fig. 7 reports, across four datasets, the numbers of inference runs that enter an over-evidence state, and measure how many of these cases are answered correctly via the Over-Evidence Monitor (OEM) as well as how many remain correct if the model continues reasoning. We find that over-evidence triggers occur more frequently after the SFT stage, but drop substantially after reinforcement learning (RL). This suggests that training on diagnosis-oriented medical data reduces excessive searching: the model is more likely to stop after gathering key evidence and rely on reasoning to produce an answer. Moreover, once OEM is triggered, the number of correct cases remains consistently low, indicating that when the model enters the over-evidence state it often becomes trapped in an incorrect medical reasoning path and repeatedly "verifies" the same mistaken hypothesis. On HLE-Med (149 questions), the trigger counts are 32/149 and 18/149, respectively, further showing that harder questions are more likely to drive the model into an over-evidence state.

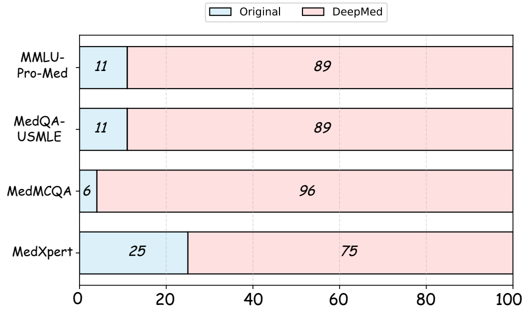

Figure 8: LLM-based comparison of whether the evidence used by the base model or DEEPMED for the final decision is more complete and reliable.

5.7 Evidence and Logical Completeness Analysis

To more directly assess how DEEPMED benefits from the DeepResearch paradigm—particularly in evidence seeking and hallucination reduction—we compare it with its base model on instances where both models answer correctly (MedXpert, MMLU-Pro, MedMCQA, and MedQA-USMLE). For each such instance, we extract the evidence chains produced by the two models and use an LLM judge to determine which chain is more comprehensive and logically reliable. The evaluation prompt is provided in Appendix. A.4. The results in Fig. 8 show that DEEPMED overwhelmingly produces evidence chains with higher completeness and reliability, suggesting that its medical judgments under the DR paradigm are more credible.

6 Conclusion

We present DEEPMED, a DeepResearch model tailored for medical questions that bridges task and tool-scaling gaps between general-purpose DR systems and medical reasoning. DEEPMED combines web-based multi-hop med-search data synthesis with Agentic SFT/RL to better couple retrieval with clinical reasoning, and introduces a difficulty-aware turn-penalty and an Over-Evidence Monitor to curb over-searching and noisy context. Experiments show consistent gains over the base model on seven medical benchmarks and improvements over several larger or more data-trained baselines. Further analysis shows that tool-augmented search–reflection–verification improves self-correction and evidence grounding, reducing hallucinations and spurious attributions.

Limitations

DeepMed demonstrates the feasibility of combining the DeepResearch paradigm with medical reasoning using a relatively small amount of training data. However, agentic training and tool-augmented rollouts incur non-trivial computational and engineering overhead, and we have not yet systematically scaled our approach to larger backbone models or substantially larger training corpora. In addition, DeepMed is evaluated only on public benchmarks and has not been validated in real clinical workflows or under regulatory standards; it should not be used as a substitute for professional medical advice or clinical decision making. Finally, our tools retrieve evidence primarily from the open web, whose availability and quality may vary over time. We hope to collaborate with medical institutions in future work to incorporate more stable and authoritative information sources and to enable more realistic evaluations.

Ethics Statement

This work does not involve human participants, patient records, or clinical studies, and thus no IRB approval is required. We rely on public benchmarks and synthesized data, and we do not collect or distribute any personally identifiable information. For web-based evidence retrieval, we only access publicly available content and report aggregated findings; we encourage compliance with website terms and applicable data policies. The system is a research prototype and must not be used as medical advice or for clinical decision making without appropriate oversight and validation.

Acknowledgements

The work is supported by the National Natural Science Foundation of China (Nos. 62272092, 62502081, 62172086), and the Fundamental Research Funds for the Central Universities under Grants (N25XQD004).

References

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024. Huatuogpt-o1, towards medical complex reasoning with llms. *Preprint*, arXiv:2412.18925.

Shan Chen, Pedro Moreira, Yuxin Xiao, Sam Schmidgall, Jeremy Warner, Hugo Aerts, Thomas Hartvigsen, Jack Gallifant, and Danielle S Bitterman. 2025. Medbrowsecomp: Benchmarking medical deep research and computer use. *arXiv preprint arXiv:2505.14963*.

Chengfeng Dou, Chong Liu, Fan Yang, Fei Li, Jiyuan Jia, Mingyang Chen, Qiang Ju, Shuai Wang, Shunya Dang, Tianpeng Li, and 1 others. 2025. Baichuan-m2: Scaling medical capability with large verifier system. *arXiv preprint arXiv:2509.02208*.

Yichun Feng, Jiawei Wang, Ruikun He, Lu Zhou, and Yixue Li. 2025. A retrieval-augmented knowledge mining method with deep thinking llms for biomedical research and clinical support. *GigaScience*, 14:giaf109.

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. *arXiv preprint arXiv:2501.12948*.

Kelly Hong, Anton Troynikov, and Jeff Huber. 2025. Context rot: How increasing input tokens impacts llm performance. Technical report, Chroma.

Xiaoke Huang, Juncheng Wu, Hui Liu, Xianfeng Tang, and Yuyin Zhou. 2025. m1: Unleash the potential of test-time scaling for medical reasoning with large language models. *arXiv preprint arXiv:2504.00869*.

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. *arXiv preprint arXiv:2412.16720*.

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. *arXiv preprint arXiv:2503.09516*.

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

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Yu He Ke, Liyuan Jin, Kabilan Elangovan, Hairil Rizal Abdullah, Nan Liu, Alex Tiong Heng Sia, Chai Rick Soh, Joshua Yi Min Tung, Jasmine Chiat Ling Ong, Chang-Fu Kuo, and 1 others. 2025. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. *npj Digital Medicine*, 8(1):187.

H. Kim, J. Sohn, A. Gilson, and 1 others. 2025. Rethinking retrieval-augmented generation for medicine: A large-scale, systematic expert evaluation and practical insights. *arXiv preprint arXiv:2511.06738*.

Ao Li, Bin Yan, Bingfeng Cai, Chenxi Li, Cunzhong Zhao, Fugen Yao, Gaoqiang Liu, Guanjun Jiang, Jian Xu, Liang Dong, and 1 others. 2025a. Quarkmed medical foundation model technical report. *arXiv preprint arXiv:2508.11894*.

Kuan Li, Zhongwang Zhang, Huifeng Yin, Rui Ye, Yida Zhao, Liwen Zhang, Litu Ou, Dingchu Zhang, Xixi Wu, Jialong Wu, and 1 others. 2025b. Websailor-v2: Bridging the chasm to proprietary agents via synthetic data and scalable reinforcement learning. *arXiv preprint arXiv:2509.13305*.

Kuan Li, Zhongwang Zhang, Huifeng Yin, Liwen Zhang, Litu Ou, Jialong Wu, Wenbiao Yin, Baixuan Li, Zhengwei Tao, Xinyu Wang, and 1 others. 2025c. Websailor: Navigating super-human reasoning for web agent. *arXiv preprint arXiv:2507.02592*.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, and 1 others. 2025a. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

Che Liu, Haozhe Wang, Jiazhen Pan, Zhongwei Wan, Yong Dai, Fangzhen Lin, Wenjia Bai, Daniel Rueckert, and Rossella Arcucci. 2025b. Beyond distillation: Pushing the limits of medical llm reasoning with minimalist rule-based rl. *arXiv preprint arXiv:2505.17952*.

Chi Liu, Derek Li, Yan Shu, Robin Chen, Derek Duan, Teng Fang, and Bryan Dai. 2025c. Fleming-r1: Toward expert-level medical reasoning via reinforcement learning. *arXiv preprint arXiv:2509.15279*.

Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, and 1 others. 2023. Benchmarking large language models on cmexam–a comprehensive chinese medical exam dataset. *arXiv preprint arXiv:2306.03030*.

Lei Liu, Xiaoyan Yang, Junchi Lei, Yue Shen, Jian Wang, Peng Wei, Zhixuan Chu, Zhan Qin, and Kui Ren. 2024. A survey on medical large language models: Technology, application, trustworthiness, and future directions. *Preprint*, arXiv:2406.03712.

K. Lu, Z. Liang, D. Pan, and 1 others. 2025a. Med-r^2: Crafting trustworthy llm physicians via retrieval and reasoning of evidence-based medicine. *arXiv preprint arXiv:2501.11885*.

Rui Lu, Zhenyu Hou, Zihan Wang, Hanchen Zhang, Xiao Liu, Yujiang Li, Shi Feng, Jie Tang, and Yuxiao Dong. 2025b. Deepdive: Advancing deep search agents with knowledge graphs and multi-turn rl. *arXiv preprint arXiv:2509.10446*.

Fnu Neha, Deepshikha Bhati, and Deepak Kumar Shukla. 2025. Retrieval-augmented generation (rag) in healthcare: A comprehensive review. *AI*, 6(9):226.

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

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, and 1 others. 2025a. Humanity's last exam. *arXiv preprint arXiv:2501.14249*.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, and 1 others. 2025b. Humanity's last exam. *Preprint*, arXiv:2501.14249.

Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Towards building multilingual language model for medicine. *Preprint*, arXiv:2402.13963.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*.

Baorong Shi, Bo Cui, Boyuan Jiang, Deli Yu, Fang Qian, Haihua Yang, Huichao Wang, Jiale Chen, Jianfei Pan, Jieqiong Cao, and 1 others. 2026. Medxiaohe: A comprehensive recipe for building medical mllms. *arXiv preprint arXiv:2602.12705*.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, and 1 others. 2025a. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.

Tongyi DeepResearch Team, Baixuan Li, Bo Zhang, Dingchu Zhang, Fei Huang, Guangyu Li, Guoxin Chen, Huifeng Yin, Jialong Wu, Jingren Zhou, and 1 others. 2025b. Tongyi deepresearch technical report. *arXiv preprint arXiv:2510.24701*.

Rahul Thapa, Qingyang Wu, Kevin Wu, Harrison Zhang, Angela Zhang, Eric Wu, Haotian Ye, Suhana Bedi, Nevin Aresh, Joseph Boen, Shriya Reddy, Ben Athiwaratkun, Shuaiwen Leon Song, and James Zou. 2025. Disentangling reasoning and knowledge in medical large language models. *Preprint*, arXiv:2505.11462.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature Medicine*, 29(8):1930–1940.

Yisu Wang, Ming Wang, Haoyuan Song, Wenjie Huang, Chaozheng Wang, Yi Xie, and Xuming Ran. 2026. Repair: Robust editing via progressive adaptive intervention and reintegration. *Preprint*, arXiv:2510.01879.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Preprint*, arXiv:2406.01574.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. 2025. Browsecomp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*.

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

Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Linhai Zhang, Yulan He, Deyu Zhou, Pengjun Xie, and 1 others. 2025a. Webwalker: Benchmarking llms in web traversal. *arXiv preprint arXiv:2501.07572*.

Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, and 1 others. 2025b. Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. *arXiv preprint arXiv:2504.00993*.

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

K. Xu, W. Hou, Y. Cheng, and 1 others. 2025. Rar^2: Retrieval-augmented medical reasoning via thought-driven retrieval. *arXiv preprint arXiv:2509.22713*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

Ailing Yu, Lan Yao, Jingnan Liu, Zhe Chen, Jiajun Yin, Yuan Wang, Xinhao Liao, Zhiling Ye, Ji Li, Yun Yue, and 1 others. 2025a. Medresearcher-r1: Expert-level medical deep researcher via a knowledge-informed trajectory synthesis framework. *arXiv preprint arXiv:2508.14880*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2025b. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, and 1 others. 2025. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*.

Wenlin Zhang, Xiaopeng Li, Yingyi Zhang, Pengyue Jia, Yichao Wang, Huifeng Guo, Yong Liu, and Xiangyu Zhao. 2025a. Deep research: A survey of autonomous research agents. *arXiv preprint arXiv:2508.12752*.

Xiaotian Zhang, Yuan Wang, Zhaopeng Feng, Ruizhe Chen, Zhijie Zhou, Yan Zhang, Hongxia Xu, Jian Wu, and Zuozhu Liu. 2025b. Med-u1: Incentivizing unified medical reasoning in llms via large-scale reinforcement learning. *arXiv preprint arXiv:2506.12307*.

Yiqun Zhang, Peng Ye, Xiaocui Yang, Shi Feng, Shufei Zhang, Lei Bai, Wanli Ouyang, and Shuyue Hu. 2025c. Nature-inspired population-based evolution of large language models. *Preprint*, arXiv:2503.01155.

Xuejiao Zhao, Siyan Liu, Su-Yin Yang, and Chunyan Miao. 2025. Medrag: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. *Preprint*, arXiv:2502.04413.

Peilin Zhou, Bruce Leon, Xiang Ying, Can Zhang, Yifan Shao, Qichen Ye, Dading Chong, Zhiling Jin, Chenxuan Xie, Meng Cao, and 1 others. 2025. Browsecomp-zh: Benchmarking web browsing ability of large language models in chinese. *arXiv preprint arXiv:2504.19314*.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *Preprint*, arXiv:2501.18362.

A Appendix

Multi-hop Med-search QA

Question:
A pediatric malignancy accounting for roughly 50% of soft tissue sarcomas in children and approximately 400 to 500 annual diagnoses in the United States comprises distinct histological categories. The most prevalent form, representing 60–70% of cases with a peak incidence in children aged 0–5 years, is typically "fusion-negative" and often driven by RAS pathway mutations.
Conversely, a more aggressive subtype, constituting 20–30% of diagnoses with a peak onset between 10 and 25 years, is characterized by the presence of chimeric transcription factors. In approximately 60% of this aggressive subtype, the pathology is driven by a specific fusion oncoprotein that is 438 amino acids in length and exhibits 10- to 100-fold greater transcriptional activity than its wild-type predecessors. This protein is distinct from a less common variant found in roughly 20% of cases.
Identify the specific reciprocal genetic rearrangement—naming the chromosomes and bands involved—that generates this primary 438-amino acid oncogenic driver.

Reasoning Path:
Childhood Rhabdomyosarcoma → Embryonal vs. Alveolar (Aggressive) → Alveolar Rhabdomyosarcoma → *PAX3::FOXO1* Fusion → **t(2;13)(q35;q14) translocation**

Answer:
t(2;13)(q35;q14) translocation

Figure 9: A case of Multi-hop Med-Search QA

A.1 A case of Multi-hop Med-Search QA

Fig. 9 shows an example of our synthesized multi-hop Med-Search QA. Notably, the question contains many numerical details. We intentionally encourage the inclusion of numerical attributes in the synthesis prompt, since numbers such as prevalence rates and medication dosages are common sources of hallucination compared to entity relations and thus typically require verification via search. This design more strongly incentivizes the model to use tools in medical contexts and promotes deeper, evidence-grounded reasoning. Such information is rarely obtainable in a comprehensive form from a single medical website, let alone from a knowledge graph. This highlights the advantage of our web-based synthesis method, which aggregates evidence across multiple authoritative sources.

A.2 Evaluation Setup

Benchmarks. We evaluate DeepMed on seven medical benchmarks as Tab. 1, we can splite them to (1) challenging dataset: **HLE-Med** and **MedXpert**, and (2) standard dataset: **MMLU-Pro-Med**, **PubMedQA**, **MedMCQA**, **MedQA**,**CMExam**.

- **HLE-Med** (Phan et al., 2025b) stands as one of the most challenging medical benchmarks, featuring questions meticulously crafted and validated by human medical experts. We evaluate the model on the `Biology/Medicine` text subset [1]with 149 questions.
- **MedXpert** (Zuo et al., 2025) is a complex benchmark designed to assess expert-level medical reasoning grounded in relevant medical literature. In our experiments, we utilize the `text` subset that contains 2,450 questions.
- **MMLU-Pro-Med** (Wang et al., 2024) provides a comprehensive evaluation of medical knowledge and reasoning capabilities. We employ the `biology` and `health` subsets, totaling 1535 questions.
- **PubMedQA** (Jin et al., 2019) contains 500 questions to evaluate the model's ability of answering clinical inquiries based on medical articles.
- **MedMCQA** (Pal et al., 2022) includes 4,183 questions designed to address real-world medical entrance exam questions.
- **MedQA** (Jin et al., 2020) consists of questions collected from professional medical examinations. We focus on the USMLE subset, comprising 1,273 questions.

[1] https://huggingface.co/datasets/futurehouse/hle-gold-bio-chem

- **CMExam** (Liu et al., 2023) consists of 6,811 questions sourced from the Chinese National Medical Licensing Examination.

Dataset	Size	Type
HLE-Med (Phan et al., 2025b)	149	MCQ + Open-ended
MedXpert (Zuo et al., 2025)	2,450	MCQ
MMLU-Pro-Med (Wang et al., 2024)	1,535	MCQ
PubMedQA (Jin et al., 2019)	500	MCQ
MedMCQA (Pal et al., 2022)	4,183	MCQ
MedQA-USMLE (Jin et al., 2020)	1,273	MCQ
CMExam (Liu et al., 2023)	6,811	MCQ

Table 4: Overview of evaluation benchmarks.

Baselines. We compare DeepMed against three categories of strong baselines. And we avoid comparing against models that rely on private databases or are not fully open-sourced (e.g., MedResearcher-R1 (Yu et al., 2025a)), as such comparisons may introduce confounding factors and bias. First, we select representative **General Large Language Models**, including Gemini-2.5-Pro, Kimi-k2-Thinking (Team et al., 2025a), DeepSeek-V3.2 (Liu et al., 2025a), and the Qwen3 series (Qwen3-32B-Thinking and Qwen3-14B) (Yang et al., 2025). Second, we evaluate state-of-the-art **Medical LLMs** that utilize diverse post-training and data synthesis strategies: HuatuoGPT-o1-70B (Chen et al., 2024) pioneers the SFT-then-RL paradigm to achieve SOTA reasoning performance; the M1 series focuses on enhancing medical reasoning via test-time scaling; MedReason-8B (Wu et al., 2025b) employs a medical knowledge graph to synthesize high-quality CoT paths; AlphaMed-70B (Liu et al., 2025b) achieves competitive results by analyzing question informativeness and applying pure RL specifically to high-informativeness questions; BaiChuan-M2-32B (Dou et al., 2025) introduces a Patient Simulator and Clinical Rubrics Generator to mimic real-world clinical decision-making, subsequently optimized via interactive GRPO; and QuarkMed-32B (Li et al., 2025a) enhances medical knowledge coverage by synthesizing vast amounts of data from medical literature and clinical records. Thirdly, we include **Deep Research Models** designed for autonomous information synthesis, such as Tongyi-DeepResearch (Team et al., 2025b) and MedResearcher-R1 (Yu et al., 2025a).

A.3 Experiments Details

Data Synthesis To synthesize multi-hop medsearch data, we first need to define suitable starting points. We initiate random walks from the 50 most common real-world drugs and diseases. During the walks, we store every visited entity description and every accessed webpage in a local cache, which helps reduce cost by avoiding redundant retrieval in subsequent visits. In total, we cached 67,024 entity descriptions and 200,584 webpages. When generating multi-hop questions, we randomly sample walk trajectories with hop lengths between 3 and 8, ensuring a more balanced distribution of question difficulty.

Training Settings Both agentic SFT and RL training are conducted using the Verl (Sheng et al., 2024) framework. The training prompt template is shown in the Fig. 12 During the agentic SFT stage, we train the model for 5 epochs and use a training batch size of 32 with a learning rate of 1×10^{-5}, applying a cosine learning rate schedule with a warm-up ratio of 0.1. The maximum sequence length is set to 40K tokens. For the agentic RL stage, we sample $G = 8$ rollouts per query, with a maximum of 30 tool calling turns per rollout. The training batch size is 64, with a PPO mini-batch size of 16. The maximum sequence length remains 40K tokens, the clipping range ϵ is set to 0.2, and the learning rate is set to 1×10^{-6}.

Inference Settings During inference, we set the temperature to 1.0 and used default values for all other parameters. We capped the maximum number of rounds at 30, and limited the OE monitor to 20 rounds.

Tool Configuration In this work, we predefine two tools to support medical reasoning: a `search` tool and a `visit` tool as MEDXIAOHE (Shi et al., 2026). The `search` tool is implemented using Serper's Google Search API[2], which retrieves web search results based on specified queries and returns both brief summaries and corresponding URLs. Jina API[3] is used for `visit` tool to get all webpage content of the specific url. Subsequently, we segment the webpage content into chunks of 100k tokens. A summarization model then processes these chunks to extract only the information relevant to the given goal, thereby preventing context overflow caused by excessively long webpages.

[2]https://serper.dev
[3]https://jina.ai

The detailed tool schemas are provided in Fig. 11. The prompt for this summarization model is shown in Fig. 13.

A.4 Prompt for Evidence Analysis

We use the prompt in Fig. 10 directly compare DEEPMED and Qwen-14B to determine which produces more well-supported and reliable answers.

A.5 A Case of DEEPMED Solving QA

Figure 14 presents a case study demonstrating how the model leverages search results to rectify prior biases and enhance medical reasoning. In this case, the model initially exhibited a strong prior bias towards the pancreatic duct (Option B) due to anatomical proximity. However, by iteratively refining its search queries to specifically investigate complications of posterior duodenal ulcers, the model retrieved critical evidence identifying the gastroduodenal artery as the structure most susceptible to erosion. This external verification allowed the model to self-correct and confidently converge on the correct diagnosis (Option D).

A.6 Medical Task Subtype Comparison

To provide a more fine-grained analysis of model capabilities, we further compare performance across medical task subtypes based on **LLM-based categorization**. These subtypes are not original human-annotated labels from the datasets; instead, they are obtained by using an LLM to categorize each question according to its content and required reasoning objective. Specifically, we group the questions into five categories: **Knowledge Recall**, **Clinical Diagnosis**, **Medical Treatment**, **Medical Management**, and **Mechanism and Pathology**.

As shown in Tables 5 and 6, **DeepMed-RL consistently improves over its base model across all subcategories**, with particularly clear gains on reasoning-intensive tasks such as **Clinical Diagnosis**, **Medical Treatment**, and **Medical Management**. In contrast, **Tongyi-DR** mainly improves **Knowledge Recall**, while showing limited or negative gains on several reasoning-related categories. These results indicate that DeepMed-RL achieves stronger and more consistent improvements on medical reasoning tasks, whereas directly applying general-domain DR methods to medicine still leaves a substantial gap.

Table 5: Performance comparison across medical task subtypes on **MedMCQA** (4,183 questions). The subtype partition is obtained via LLM-based categorization.

Model	Knowledge Recall(3026)	Clinical Diagnosis(440)	Treatment (246)	Management (342)	Mechanism & Pathology(129)	Overall
Qwen3-30BA3B	71.91	79.55	76.42	71.05	78.29	73.11
Tongyi-DR	77.26	73.18	69.92	64.91	71.32	75.21
Qwen3-14B	66.89	73.86	72.36	66.37	75.97	68.18
DeepMed-RL	**83.18**	**84.77**	**82.52**	**76.02**	**79.07**	**82.60**

Table 6: Performance comparison across medical task subtypes on **MedQA-USMLE** (1,273 questions). The subtype partition is obtained via LLM-based categorization.

Model	Knowledge Recall(3026)	Clinical Diagnosis(440)	Treatment (246)	Management (342)	Mechanism & Pathology(129)	Overall
Qwen3-30BA3B	91.57	87.87	92.57	80.80	92.83	88.14
Tongyi-DR-30BA3B	**93.98**	87.52	91.22	80.40	92.83	87.90
Qwen3-14B	89.16	81.20	85.81	73.60	89.24	82.17
DeepMed-RL-14B	91.57	**88.05**	**93.92**	**81.20**	91.48	**88.22**

A.7 Medical Knowledge Coverage Analysis

To further examine the coverage of the synthesized medical knowledge, we analyze the nodes collected during our multi-hop question generation process from two complementary perspectives: (i) department-based classification and (ii) entity ontology-based classification. In total, our synthesized data covers 11,191 unique medical entities. Since the question generation process essentially performs random walks over web-scale medical knowledge, with starting points initialized from diverse medical departments, the resulting entities achieve broad coverage across both specialties and concept types.

Table 7 shows the department-based distribution of the extracted entities. The synthesized data spans more than 20 medical specialties, including **Basic Medical Science**, **Neurology**, **Infectious Disease**, **Cardiovascular**, and **Oncology**, indicating strong breadth across both foundational and clinical domains.

Table 8 presents the ontology-based distribution of the same entity set. The synthesized data covers more than **12 ontology types**, including **Disease/Condition**, **Protein/Enzyme/Receptor**, **Drug/Medication**, **Anatomical Structure**, **Pathway/Process**, and **Diagnostic Test**, demonstrating substantial diversity in medical knowledge representation.

Overall, the synthesized data exhibits strong coverage in both breadth and diversity. **Disease/Condition** and **Symptom/Sign** account for a large proportion of the entities, while categories such as **Drug/Medication** and **Protein/Enzyme/Receptor** also have substantial coverage. These statistics suggest that our synthesized data provides broad and representative support for medical reasoning and question construction.

Table 7: Department-based classification of medical entities extracted from the synthesized data. The total number of unique entities is 11,191.

Department	**Count**	**Department**	**Count**
Basic Medical Science	1,425	Immunology	590
Neurology	1,131	OBGYN/Reproductive	525
Infectious Disease	1,032	Hematology	505
Cardiovascular	909	Cross-System	489
Oncology	796	Respiratory	447
Gastrointestinal	647	Endocrine	425
Musculoskeletal	618	Renal/Urinary	407
Psychiatry	384	Dermatology	293
Ophthalmology	225	Pediatrics	121
Otolaryngology	94	Other	128

Table 8: Ontology-based classification of medical entities extracted from the synthesized data. The total number of unique entities is 11,191.

Ontology Type	**Count**	**Ontology Type**	**Count**
Disease/Condition	4,578	Biochemical Substance	498
Protein/Enzyme/Receptor	1,207	Gene/Genetic	436
Drug/Medication	1,117	Symptom/Sign	428
Anatomical Structure	767	Pathogen	294
Pathway/Process	604	Surgical Procedure	289
Clinical Concept	599	Diagnostic Test	287
Other	87		

A.8 Comparison with Existing Agent and RAG Frameworks

To our knowledge, **DeepMed** is the first work that explicitly investigates the gap that arises when applying the *DeepResearch* paradigm to medical problems. Accordingly, for the comparison with deep research agents, we include **Tongyi-DR-30BA3B**, a strong open-source deep research model under 200B parameters that originally motivated our study by revealing the limitations of general-domain deep research on medical tasks. As an additional baseline, we further include **DeepDive-32B** (Lu et al., 2025b). The results are shown in Table 9. Overall, **DeepMed-RL-14B** achieves competitive or superior performance across the evaluated medical benchmarks, despite using a smaller base model.

We further compare **DeepMed-RL** with recent medical RAG methods from 2025, including **Med-R^2** (Lu et al., 2025a), **RAR^2** (Xu et al., 2025), and **MedCPT+Filtering+Query-Reform** (Kim et al., 2025), on their shared evaluation benchmarks. For each baseline, we report the best result stated in the corresponding paper. As shown in Table 10, **DeepMed-RL** shows a clear advantage on several representative datasets, including **MedQA-USMLE**, **MedMCQA**, **PubMedQA**, and **MedXpertQA**. These results suggest that our approach compares favorably not only against general deep research agents, but also against recent retrieval-augmented medical reasoning systems.

A.9 Detail Analysis of Web Search Quality

The DeepResearch paradigm is not in conflict with professional medical databases; rather, the two are complementary. In practice, the retrieval component can be directly connected to professional medical databases. At the same time, web-scale search offers broader accessibility and greater flexibility, while many professional medical databases are proprietary and less adaptable in open settings. Notably, **DeepMed** achieves stronger performance than competing baselines despite using the smallest parameter count and minimal training data, which provides empirical evidence for the reliability of web-based medical information when coupled with effective filtering and reasoning.

In Section 5.5, we analyze the distribution of websites visited by DeepMed and find that more than **78.4%** of visited sources are authoritative medical websites, such as **FDA**, **PubMed**, and **CDC.gov**, as verified by our medical team. Among the remaining sources, more than **16%** come from professional knowledge bases such as **Wikipedia**. These results indicate that the model naturally tends to rely on high-quality medical sources during reasoning.

Like some other fields of work (Zhang et al., 2025c; Wang et al., 2026), we compare the category distribution of raw search engine returns with that of the websites actually visited by the model, as shown in Table 11. The results show that while **24.5%** of search returns fall into the **Other** category, this proportion is reduced to only **4.6%** in the model's actual visits. In contrast, authoritative categories such as **Medical Literature**, **Clinical Reference**, and **Professional Knowledge Bases** all occupy a larger proportion in the final visited pages than in the initial search results. This suggests that, after training, DeepMed acquires an effective information filtering capability: it preferentially selects authoritative medical literature and clinical references, while filtering out most noisy or less relevant content.

These findings indicate that although web data

Table 9: Comparison with deep research agent baselines.

Model	HLE-Med	MedXpert	MMLU-Pro-Med	PubMedQA	MedMCQA	MedQA-USMLE	CMExam
Tongyi-DR-30BA3B	28.19	35.40	79.54	76.20	75.21	87.90	87.59
DeepDive-32B (Lu et al., 2025b)	21.77	32.82	76.87	78.60	73.87	85.39	89.24
DeepMed-RL-14B	26.84	**36.14**	**79.93**	**81.80**	**82.60**	**88.22**	**91.19**

Table 10: Comparison with recent medical RAG methods. For each baseline, we report the best number stated in the corresponding paper.

Method	Base Model	MedQA-USMLE	MedMCQA	PubMedQA	MedXpertQA
Med-R^2 (Lu et al., 2025a)	LLaMA3.1-70B	86.37	73.36	78.24	26.84
RAR^2 (w/o train) (Xu et al., 2025)	Qwen2.5-7B-Instruct	64.89	63.71	—	17.59
MedCPT+Filtering+Query-Reform (Kim et al., 2025)	LLaMA-3.1-8B	76.2	—	64.8	21.8
DeepMed-RL	Qwen3-14B	**88.22**	**82.60**	**81.80**	**36.14**

Table 11: Comparison between the category distribution of raw search engine returns and the websites actually visited by DeepMed.

Category	Search (%)	Visit (%)
Medical Literature	23.6	29.2
Clinical Reference	19.2	24.0
Professional Knowledge Bases	13.8	17.0
Medical Education	9.7	13.4
Public Health	5.4	7.1
Community	3.8	4.7
Other	24.5	4.6

inevitably contains noise, the trained DeepMed is able to identify and utilize reliable medical information effectively. More importantly, the *DeepResearch* paradigm is not inherently tied to open web search, and can be seamlessly extended to incorporate professional medical databases when such resources are available.

Prompt for Evidence Analysis

You are an expert evaluator for medical question answering systems. Your task is to judge which model provides **MORE COMPLETE EVIDENCE** to support its answer.

Question
{question}

Model A (Original Model) – Single-turn reasoning
{original_evidence}

Model B (RL Model) – Multi-turn with web search
{rl_evidence}

Evaluation Criteria
Consider the following aspects when judging evidence completeness:
1. **Breadth of Sources**: Does the model cite multiple authoritative sources (medical textbooks, clinical guidelines, research papers)?
2. **Depth of Explanation**: Does the model explain the underlying medical mechanisms, not just state facts?
3. **Clinical Relevance**: Does the model provide clinically relevant details (dosages, contraindications, diagnostic criteria)?
4. **Verification**: Does the model verify its claims with external evidence?
5. **Reasoning Chain**: Is the logical chain from evidence to conclusion clear and complete?

Instructions
- Model A uses internal knowledge with detailed reasoning in a single turn.
- Model B uses web search to find and cite external sources across multiple turns.
- Judge based on **EVIDENCE COMPLETENESS**, not just answer correctness (both got the correct answer).
- Consider both quality and quantity of supporting evidence.

Response Format
Respond with a JSON object:
{ "winner": "A" or "B", "reason": "Brief explanation (1-2 sentences)" }
Only output the JSON, nothing else.

Figure 10: Prompt for evidence Analysis.

Function Calling Schema

```
// Tool 1: Search Tool
{
    "type": "function",
    "function": {
      "name": "search",
      "description": "Perform Google web searches and
     return top results.",
      "parameters": {
        "type": "object",
        "properties": {
          "query": {
            "type": "array",
            "items": { "type": "string" },
            "description": "Array of query strings
     (1-5 queries).",
            "minItems": 1,
            "maxItems": 5
          }
        },
        "required": [ "query" ]
      }
    }
},
// Tool 2: Visit Tool
{
    "type": "function",
    "function": {
      "name": "visit",
      "description": "Visit webpage(s) and return
     summary.",
      "parameters": {
        "type": "object",
        "properties": {
          "url": {
            "type": "array",
            "items": { "type": "string" },
            "description": "The URL(s) to visit (1-3
     URLs).",
            "minItems": 1,
            "maxItems": 3
          },
          "goal": { "type": "string" }
        },
        "required": [ "url" ]
      }
    }
}
```

Figure 11: JSON schema of the tools used.

System Prompt

You are a Medical deep research assistant. Your core function is to conduct thorough, multi-source investigations into any topic. You must handle both broad, open-domain inquiries and queries within specialized academic fields. For every request, synthesize information from credible, diverse sources to deliver a comprehensive, accurate, and objective response. When you have gathered sufficient information and are ready to provide the definitive response, you must enclose the entire final answer within <answer></answer> tags.

Figure 12: System prompt for training.

Prompt for summary model

Please process the following webpage content and user goal to extract relevant information:

Webpage Content
{content}

User Goal
{goal}

Task Guidelines

1. **Content Scanning for Rationale**: Locate the **specific sections/data** directly related to the user's goal within the webpage content.
2. **Key Extraction for Evidence**: Identify and extract the **most relevant information**. Ensure you do not miss any key details. Output the **full original context** as much as possible (it can exceed three paragraphs).
3. **Summary Output for Summary**: Organize into a concise paragraph with logical flow, prioritizing clarity, and assess the contribution of the information to the goal.

Output Format: JSON format containing "rational", "evidence", and "summary" fields.

Figure 13: Prompt used for the summary model.

Case Study: Self-Correction from Search Results

User: A 55-year-old man comes to the physician because of a 3-week history of intermittent burning epigastric pain. His pain improves with antacid use and eating, but returns approximately 2 hours following meals. He has a history of chronic osteoarthritis and takes ibuprofen daily. Upper endoscopy shows a deep ulcer located **on the posterior wall of the duodenal bulb**. This ulcer is most likely to erode into which of the following structures? (A) Splenic vein (B) Pancreatic duct (C) Descending aorta (D) Gastroduodenal artery (E) Transverse colon

---

<think>
We have a medical question: a duodenal ulcer on the posterior wall of the duodenal bulb. Which structure is most likely to be eroded? We need to recall ... The pancreatic duct runs in the posterior wall of the duodenum. **So the most likely answer is B: pancreatic duct.** But let's verify with resources. Use search.
</think> **[Comment]** The model starts with a strong prior bias towards Option B due to anatomical proximity, ignoring the vascular option initially.
<tool_call>
{ "name": "search", "arguments": { "query": ["posterior wall duodenal ulcer most likely to erode into pancreatic duct", "duodenal ulcer complication pancreatic duct erosion"] } }
</tool_call>
<tool_response>
...
(Result 9) Title: Hepatic arterial haemorrhage caused by duodenal ulcer Abstract: by S Caragher · 2022 Most posterior wall duodenal bulb ulcers erode into the **gastroduodenal artery** ... url: https://pmc.ncbi.nlm.nih.gov/articles/PMC9628537/.
</tool_response>

- -

[Comment] The model is verifying the answer... **(13 turns)**

- -

<think>
The search results suggest that erosion into the **gastroduodenal artery** is a major complication of posterior ulcers, more so than direct pancreatic duct erosion. I need to confirm the anatomical position of the GDA relative to the posterior duodenal wall to be certain.
</think> **[Comment]** The model realized that it might be wrong and further explored the details.
<tool_call>
{ "name": "visit", "arguments": { "goal": "Find mention of specific structures eroded by posterior duodenal ulcer.", "url": ["https://radiopaedia.org/cases/upper-gi-bleeding-duodenal-ulcer"] } }
</tool_call>
<tool_response>
... Case Discussion: The **gastroduodenal artery** is located just behind the posterior wall of the first part of the duodenum. Consequently, ulcers in this location may erode into the artery causing severe haemorrhage ...
<tool_response>

<think>
The duodenal bulb (first part) is closely related to the **gastroduodenal artery** posteriorly. A deep ulcer on the posterior wall is classically known to erode into this artery, causing life-threatening hemorrhage. While the pancreas is also posterior, the GDA is the structure most susceptible to high-pressure erosion in this specific context. Thus answer: D. We need to call the finish tool.
</think>
<Answer> (D) Gastroduodenal artery **</Answer>**

Figure 14: **Example of search-assisted medical reasoning.** The figure shows the model's process of correcting its initial incorrect intuition (Option B) by retrieving external evidence regarding the gastroduodenal artery, leading to the correct diagnosis (Option D).