DEEPMED Search: An Open-Source Agentic Platform for Medical Deep Research with Introspective Verification

Maolin Liu[1] , **Fanyu Xu**[1] , **Ruoqing Xu**[1] , **Jiahang Zhang**[1] , **Hao Wang**[1,*] and **Rui Wang**[2]
[1]School of Computer Science and Software Engineering, Shanghai University
[2]Department of Computer Science and Engineering, Shanghai Jiao Tong University
{23721623, 24122195, 50x20, kaito1412, wang-hao}@shu.edu.cn, wangrui12@sjtu.edu.cn

Abstract

Navigating the deluge of heterogeneous medical data, from academic literature (PubMed) to clinical guidelines (Web) and private knowledge bases remains a critical bottleneck for evidence-based medicine. While commercial black-box tools lack transparency, standard open-source RAG implementations frequently suffer from "reasoning drift" when handling complex, long-tail queries. We present **DEEPMED SEARCH**, a fully open-source, agentic platform designed for transparent medical deep research. Built on a high-performance NEXT.JS architecture, DEEPMED SEARCH features a source-adaptive router that autonomously dispatches sub-queries to PubMed, web search, or local graph-based knowledge bases based on information density. Crucially, the platform integrates an introspective verification module, powered by a causal-consistent multi-agent debate framework, to validate retrieved evidence against diagnostic logic before synthesis. To demonstrate its robustness, we showcase DEEPMED SEARCH's ability to autonomously decompose high-difficulty rare disease queries, filter out confounding noise, and generate structured, citation-backed research reports in minutes. By open-sourcing this software, we provide the community with a robust infrastructure to democratize access to trustworthy, glass-box medical reasoning in research and prototyping settings, which is publicly available at: https://www.deepmedsearch.cloud and the demonstration video is available at: https://youtu.be/4U4aok8yLpk.

1 Introduction

Large Language Models (LLMs) have revolutionized clinical question answering [Singhal *et al.*, 2025; Tu *et al.*, 2023; Yang *et al.*, 2022], yet their deployment in high-stakes evidence-based medicine remains hindered by limited transparency and auditability [Ji *et al.*, 2023; Rudin, 2019]. Medical deep research requires synthesizing heterogeneous data from academic literature, constantly updating clinical guidelines, and structured patient records. While Retrieval-Augmented Generation (RAG) is widely adopted to ground model outputs [Lewis *et al.*, 2020; Karpukhin *et al.*, 2020; Izacard and Grave, 2021], standard RAG implementations often rely on static, linear pipelines that are vulnerable to *Retrieval-Induced Reasoning Drift*, especially in complex, long-tail scenarios such as rare disease diagnosis [Mallen *et al.*, 2023]. In these contexts, standard semantic retrievers suffer from commonality bias [Kandpal *et al.*, 2023], preferentially surfacing generic evidence that supports common mimic conditions and causing the LLM to be hijacked by plausible but non-discriminative distractors. Although commercial AI research tools attempt to address this issue, their black-box nature prevents clinicians from auditing the causal logic behind retrieved evidence, which remains a non-negotiable requirement in clinical workflows.

To democratize trustworthy and auditable AI in healthcare, we introduce DEEPMED SEARCH, a fully open-source, agentic platform engineered for transparent medical deep research. Moving beyond simple RAG frameworks, DEEPMED SEARCH introduces a source-adaptive router that autonomously orchestrates sub-queries across heterogeneous endpoints, including PubMed APIs, live web search, and a local Neo4j-based knowledge graph. Furthermore, to ensure the causal alignment and faithfulness of the retrieved evidence, the platform features an introspective verification module [Kiciman *et al.*, 2024]. This engine employs contrastive hypothesis expansion to filter confounding noise and utilizes a multi-agent adversarial debate framework [Du *et al.*, 2024; Yao *et al.*, 2022] to rigorously stress-test reasoning paths before final synthesis.

In this paper, we showcase the system's glass-box interactivity, while clinicians can observe DEEPMED SEARCH as it autonomously decomposes complex queries, routes searches, and conducts real-time multi-agent deliberations. Ultimately, the system generates structured, citation-backed research reports while maintaining complete interpretability [Xiong *et al.*, 2024].

2 System Architecture

DEEPMED SEARCH is engineered as a modular, agentic workflow designed to transform high-level clinical inquiries into structured research insights. As illustrated in Figure 1, the platform transitions from autonomous data orchestration

*Corresponding Author.

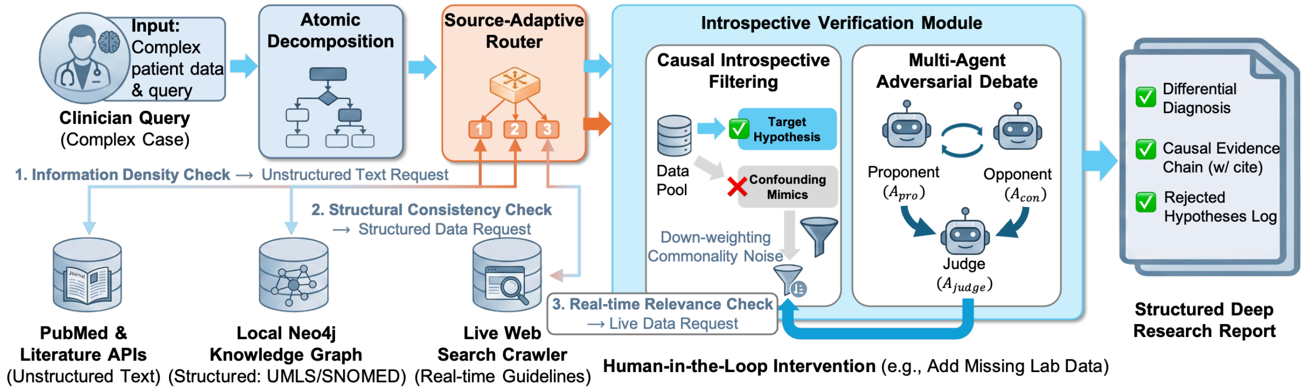

Figure 1: System architecture of DEEPMED SEARCH. The platform integrates a source-adaptive router for multi-endpoint evidence retrieval, causal introspective filtering to mitigate commonality bias, and a multi-agent debate engine to verify reasoning and synthesize reports.

to rigorous causal verification across three primary stages: Source-Adaptive Routing, Causal Introspective Filtering, and Multi-Agent Synthesis.

2.1 Source-Adaptive Routing

A core pillar of DEEPMED SEARCH is its ability to navigate heterogeneous data landscapes. Unlike monolithic RAG systems, our platform incorporates a **Source-Adaptive Router** that treats the information retrieval process as a sequential decision-making task.

Upon receiving a clinical query, the system first performs *Atomic Decomposition*, breaking the input into distinct clinical constraints (e.g., symptomatic triad, temporal progression, and treatment resistance). The Router then dynamically evaluates the information density of each sub-query to dispatch them to optimal endpoints:

- **Scientific Evidence:** Direct API integration with **PubMed** for fetching peer-reviewed literature and clinical trial abstracts [Qiu *et al.*, 2025].
- **Pathophysiological Logic:** Querying a local NEO4J knowledge graph (integrating UMLS and SNOMED-CT) to identify multi-hop causal links between symptoms [Pearl, 2009].
- **Real-time Guidelines:** Utilizing a **Web Search** crawler to retrieve the latest consensus statements and medical news.

2.2 Causal Introspective Filtering

To mitigate the *Commonality Bias* identified in clinical LLMs [Kandpal *et al.*, 2023], the platform utilizes a *Causal Introspective Filtering* mechanism. This stage aligns the retrieved evidence with diagnostic logic via two steps: **1) Contrastive Hypothesis Expansion:** The system proactively formulates "confounding hypotheses", prevalent diseases that mimic the target long-tail condition (e.g., generating "Essential Hypertension" when analyzing potential "Pheochromocytoma") [Park and Lee, 2024]. **2) Evidence Down-weighting:** Chunks are filtered based on a discriminative score $S_i = \frac{sim(c_i, H_{target})}{sim(c_i, H_{mimic}) + \epsilon}$. Here, $\text{sim}(\cdot, \cdot)$ denotes the cosine similarity between BGE-M3 sentence embeddings, and ϵ is set to 10^{-6} to avoid division by zero. A candidate chunk is regarded as discriminative when S_i is larger than the threshold τ_s; chunks failing to meet a dynamic threshold are tagged as "non-discriminative noise" and visually dimmed in the interface, ensuring the model's attention remains on the long-tail signals.

2.3 Multi-Agent Synthesis

The final reasoning phase is handled by an introspective verification module, instantiated as a multi-agent adversarial debate framework [Chan *et al.*, 2023]. The system deploys three distinct agents: a **Proponent** (A_{pro}) that constructs reasoning chains from evidence to diagnosis, an **Opponent** (A_{con}) that identifies missing causal links or inconsistencies, and a **Judge** (A_{judge}) that delivers the final verdict once reasoning converges (Information Gain $< \tau$).

Crucially, rather than producing a single-line answer, the synthesis engine generates a structured research report. These reports are citation-backed and organized into sections: Differential Diagnosis, Supporting Evidence (from both Graph and Text), and Rejected Hypotheses. This "glass-box" output allows clinicians to audit the platform's logic in real-time.

3 Implementation Details

The DEEPMED SEARCH backend is engineered for high-throughput concurrent processing to support real-time clinical deep research workflows. **Frontend & Orchestration:** Built on NEXT.JS, the frontend integrates a *WebSocket*-driven UI for streaming real-time agent deliberations. A dedicated orchestration layer handles the concurrent dispatch of sub-queries via the Source-Adaptive Router. **Knowledge Engines:** Structured multi-hop reasoning is powered by a NEO4J graph database integrating UMLS and SNOMED-CT. For unstructured evidence, we use MILVUS to index the MIRAGE corpus (Textbooks and PubMed) [Xiong *et al.*, 2024], utilizing BGE-M3 for dense vector similarity coupled with a lightweight cross-encoder for precise re-ranking. **LLM Serving:** To meet the computational demands of deploying three concurrent agents (Proponent, Opponent, Judge), we serve models like QWEN2.5-14B and DEEPSEEK-V3 via VLLM,

Query: *"Patient presents with paroxysmal hypertension, headache, palpitations, and sweating. Standard anti-hypertensive drugs ineffective."*

× **Vanilla RAG Response:**
Diagnosis: **Essential Hypertension**. Treatment involves lifestyle changes...
[Error: Fixates on common disease; ignores paroxysmal & sweating cues]

✓ **DEEPMED SEARCH (Ours):**
Diagnosis: **Pheochromocytoma**.
Reasoning: 1. *Contrastive Filter:* 'Essential Hypertension' downweighted (cannot explain sweating). 2. *Debate Verdict:* Symptom triad strongly indicates catecholamine-secreting tumor.

Table 1: Case study: DEEPMED SEARCH actively filters common disease noise (Hypertension) to correctly identify the rare condition.

leveraging *PagedAttention* to optimize memory footprint and maximize throughput. **Report Generation Engine:** A custom synthesis module compiles the verified evidence into a structured markdown/PDF report, automatically mapping reasoning steps to specific reference nodes (PubMed IDs or Graph edges) to ensure full traceability.

4 Demonstration Scenarios

We present using real-world rare disease cases from the MedR-Bench dataset, as exemplified in Table 1. The web interface allows clinicians to input complex case reports and interactively audit the deep research process.

4.1 Scenario 1: Autonomous Routing & Causal Filtering

Consider a patient presenting with episodic hypertension, headaches, and palpitations. A standard RAG system typically retrieves generic guidelines for "Primary Hypertension", leading to reasoning drift. In DEEPMED SEARCH: **1) Routing in Action:** The Router identifies the temporal cue ("episodic") and dispatches parallel queries to Neo4j and PubMed. **2) Noise Filtering:** The "Evidence Panel" marks generic hypertension evidence with a *Low Relevance* tag (visualized as dimmed text), explicitly stating it was downweighted due to overlap with common confounders. Conversely, evidence pointing to *Pheochromocytoma* is dynamically highlighted.

4.2 Scenario 2: Glass-Box Debate & Report

For complex diagnoses, the clinician can activate the introspective debate mode. **1) Real-Time Deliberation:** As illustrated in Figure 2, a streaming panel opens where A_{pro} argues for Pheochromocytoma citing paroxysmal symptoms, while A_{con} challenges the evidence density. **2) Human-in-the-Loop:** Crucially, the clinician can inject missing parameters (e.g., *"Add lab result: elevated metanephrines"*) directly into the debate stream. **Structured Output:** The A_{judge} finalizes the verdict, and the system instantly compiles a deep research report, linking the diagnosis to verified graph nodes.

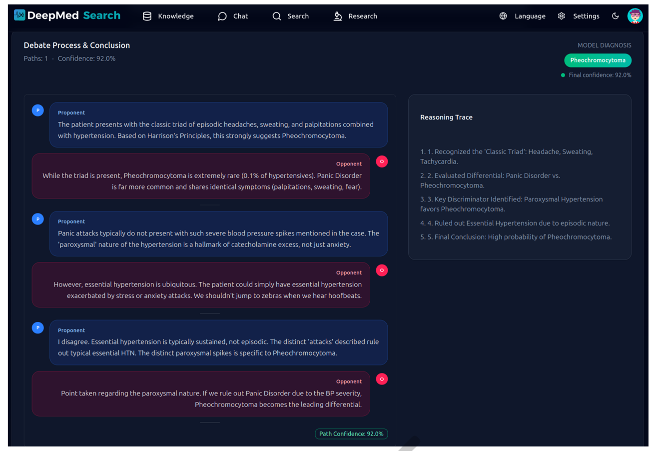

Figure 2: The Multi-Agent Debate Panel. Users can monitor agents as they argue over the diagnosis, providing transparent reasoning logic.

5 Evaluation

As summarized in Table 2, DEEPMED uniquely integrates source-adaptive routing and multi-agent verification, bridging the gap between vanilla RAG and commercial black-box tools.

Performance Gains: We evaluated on the MIRAGE benchmark [Xiong *et al.*, 2024]. DeepMed Search consistently outperforms standard RAG baselines. On the challenging PubMedQA subset, our method improves accuracy by approximately 9.0% over vanilla RAG (using Qwen3-8B). For rare diseases, experiments on the MedR-Bench dataset show that our approach achieves an accuracy of 76.43% (using DeepSeek-v3.2), bridging the gap between automated systems and expert-level reasoning.

Capabilities	Vanilla RAG	Commercial	Ours
Open-Source Transparency	✓	×	✓
Adaptive Routing (Web/Graph)	×	✓	✓
Causal-Consistent Filtering	×	×	✓
Multi-Agent Verification	×	×	✓
Structured Deep Reports	×	✓	✓

Table 2: Feature comparison. DEEPMED SEARCH bridges glass-box transparency with practical research-oriented system capabilities.

6 Conclusion

DEEPMED SEARCH provides a transparent, agentic platform for medical deep research. It combines source-adaptive routing, causal-consistent filtering, and multi-agent verification to mitigate retrieval-induced reasoning drift in medical LLMs [Park and Lee, 2024] and generate structured, citation-backed reports. Current limitations include additional computational and latency overhead and the lack of a formal clinician user study. Future work will improve efficiency, involve clinicians in evaluation, and extend the framework to multimodal data such as medical imaging [Wang *et al.*, 2025].

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 62306173.

References

[Chan *et al.*, 2023] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*, 2023.

[Du *et al.*, 2024] Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first international conference on machine learning*, 2024.

[Izacard and Grave, 2021] Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: main volume*, pages 874–880, 2021.

[Ji *et al.*, 2023] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, March 2023.

[Kandpal *et al.*, 2023] Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 15696–15707. PMLR, 2023.

[Karpukhin *et al.*, 2020] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 6769–6781, 2020.

[Kiciman *et al.*, 2024] Emre Kiciman, Robert Osazuwa Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research (TMLR)*, August 2024. Selected for presentation at ICLR 2025.

[Lewis *et al.*, 2020] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[Mallen *et al.*, 2023] Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 9802–9822, 2023.

[Park and Lee, 2024] Seong-Il Park and Jay-Yoon Lee. Toward robust ralms: Revealing the impact of imperfect retrieval on retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 12:1686–1702, 2024.

[Pearl, 2009] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, 2nd edition, 2009.

[Qiu *et al.*, 2025] Pengcheng Qiu, Chaoyi Wu, Shuyu Liu, Weike Zhao, Zhuoxia Chen, Hongfei Gu, Chuanjin Peng, Ya Zhang, Yanfeng Wang, and Weidi Xie. Quantifying the reasoning abilities of llms on real-world clinical cases. *arXiv preprint arXiv:2503.04691*, 2025.

[Rudin, 2019] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

[Singhal *et al.*, 2025] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature medicine*, 31(3):943–950, 2025.

[Tu *et al.*, 2023] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, Basil Mustafa, Aakanksha Chowdhery, Yun Liu, Simon Kornblith, David Fleet, Philip Mansfield, Sushant Prakash, Renee Wong, Sunny Virmani, Christopher Semturs, S Sara Mahdavi, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Karan Singhal, Pete Florence, Alan Karthikesalingam, and Vivek Natarajan. Towards generalist biomedical ai, 2023.

[Wang *et al.*, 2025] Ziyue Wang, Junde Wu, Linghan Cai, Chang Han Low, Xihong Yang, Qiaxuan Li, and Yueming Jin. Medagent-pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow. *arXiv preprint arXiv:2503.18968*, 2025.

[Xiong *et al.*, 2024] Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics ACL 2024*, pages 6233–6251, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics.

[Yang *et al.*, 2022] Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, Christopher A Harle, Gloria Lipori, Duane A Mitchell, William R Hogan, Elizabeth A Shenkman, Jiang Bian, and Yonghui Wu. Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records, 2022.

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