MedBrowseComp: Benchmarking Medical Deep Research and Computer Use

Anonymous Author(s)
Affiliation
Address
email

Abstract

Large language models (LLMs) are increasingly envisioned as decision-support tools in clinical practice, yet safe clinical reasoning demands the integration of heterogeneous knowledge bases—trials, primary studies, regulatory documents, and cost data—under strict accuracy constraints. Existing evaluations typically rely on synthetic prompts, reduce the task to single-hop factoid queries, or conflate reasoning with open-ended text generation, leaving their real-world utility unclear. To close this gap, we present **MedBrowseComp**, the first benchmark that systematically tests an agent's ability to reliably retrieve and synthesize multi-hop medical facts from up-to-date, domain-specific knowledge bases. MedBrowseComp holds 1,000+ human-curated questions that mirror clinical scenarios in which practitioners must reconcile information fragmented over many sources that are potentially conflicting. Applying MedBrowseComp to frontier agentic systems reveals **marked performance shortfalls as low as 10%**. MedBrowseComp reveals critical gaps between current LLM performance and clinical usage, providing a testbed to guide future model and toolchain improvements for reliable medical information seeking.

1 Introduction

LLMs have saturated static knowledge benchmarks, diminishing their utility for advancing the field [1–6]. This exposes an evaluation gap: Legacy benchmarks test static recall while agentic systems should plan, browse, and synthesize real-time evidence [7–9]. The progression from chatbots to autonomous agents promises access to real-time data and complex information gathering previously exclusive to humans [10–12]. Web-enabled agents could retrieve any well-specified fact from the open web, even across thousands of pages [13–15]. This is especially compelling in medicine, where clinical decisions require integrating current information from journal articles, trial registries, guidelines, and drug databases [16–22]. However, the community lacks unified benchmarks for evaluating complex medical retrieval at scale.

Current LLMs frequently hallucinate, generating confident but incorrect statements [23]. In high-stakes medicine, these errors can misinform clinicians and erode trust. Benchmarks must test reasoning, navigation, and evidence grounding [18, 19, 23]. Popular medical benchmarks (MMLU, MedQA, WorldMedQA) test memorizable information, with frontier models achieving near-ceiling scores [3, 24, 25]. Yet, they sidestep real-world hurdles like pagination, obsolete links, and contradictory evidence [26, 27].

Medicine requires integrating scattered information across heterogeneous sites. New benchmarks must force agents to conduct multi-hop, evidence-grounded searches to: **1)** Measure real-world navigation and information reconciliation capabilities, and **2)** Dynamically stress-test systems as evidence evolves. We introduce MedBrowseComp, evaluating AI agents' complex medical information retrieval via web browsing. Inspired by BrowseComp [28], it focuses on short, objective, verifiable answers. **We designed collaboratively with physicians using HemOnc.org, one of the largest**

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Computer Use Agent Questions

Base Question:
"On Hemonc.org, find/search the clinical trial id that best describes the efficacy of Lenalidomide monotherapy compared to Erythropoietin beta and Lenalidomide when used to treat Myelodysplastic syndrome."
Output format: NCT

Extended Questions:
- Track pubmed id for the trial
- Track given trial start date
- Track second author of the pubmed paper link to trial

Website Ground Truth
Link-https://hemonc.org/wiki/Myelodysplastic_syndrome

Additional Context
Follow-up queries expand on the clinical trial data to trace connections between authors, publications, and timeline information.

Deep Research Questions

Hop 1: Clinical Trial Ingredient
"For clinical trial NCT00843882, among the more effective regimen ingredients, find which ingredient starts with 'L'."
INGREDIENT: LENALIDOMIDE

Hop 2: Company with Latest FDA Approval
"Then, find which company has the latest FDA approval date up till Dec, 2024 for this identified ingredient."
COMPANY: CELGENE CORPORATION (BRISTOL MYERS SQUIBB)

Hop 3: Patent Expiration Date
"Then, for this identified ingredient that was last approved up till Dec, 2024, when is its patent expiration date?"
DATE: Apr 27, 2027

Hop 4: FDA Exclusivity Date
"Then, for this identified ingredient that was last approved up till Dec, 2024, when is its exclusivity date according to the FDA?"
DATE: 5/28/2026

Hop 5: Stock Market Data
"If this company is listed on any US stock market, provide: 1. The stock ticker symbol 2. The opening stock price on the FDA approval date"
TICKER: BMY (Bristol Myers Squibb) OPENING PRICE: $46.73

Figure 1: Example question constructions for MedBrowseComp.

structured medical wikis maintained by oncologists for 6 years. State-of-the-art systems achieve <50% overall accuracy, and <10% on hardest questions.

The primary contributions of this work are: **The MedBrowseComp Dataset:** Novel medical fact-seeking questions requiring web browsing with short, easy **verifiable** answers. We are also a pioneer in comprehensive benchmarking using linked domain knowledge. **Baseline Performance Analysis:** Empirical evaluation of state-of-the-art LLMs and agentic systems, establishing initial benchmarks. **Demonstration of Capability Gaps:** Evidence of gaps between general browsing agents and specialized medical information-seeking skills.

2 Related Work and Our Methods

GAIA pioneered AI evaluation for tool use and web browsing, combining multi-modal input, reasoning, and external tools [29]. WebWalker introduced dual-agent frameworks for horizontal browsing and vertical site navigation, with WebWalkerQA testing multi-hop questions across complex hyperlink structures [30, 27]. FRAMES evaluated RAG systems for factual correctness and reasoning [31], while SimpleQA targeted LLM hallucinations but remained solvable with basic searches [32, 9].

Humanity's Last Exam tests expert knowledge through specialized questions targeting model gaps across disciplines [6]. BrowseComp and BrowseComp-ZH challenge agents with hard-to-find facts using reverse-engineered questions [28, 33]. Leading models achieve <10% accuracy on BrowseComp's 1,255 English questions and 10-20% on BrowseComp-ZH's Chinese tasks, revealing persistent retrieval limitations [34].

MedBrowseComp fills this gap with medical specialization, featuring 1000+ questions (605 deep research, 484 computer use[1]) requiring exploration of reputable sources, terminology interpretation, and evidence-based reasoning. Its high-stakes domain exposes the limitations of generic benchmarks while offering an easy expandable, and updatable design for evolving medical knowledge.

We constructed MedBrowseComp using HemOnc.org, the largest freely available hematology/oncology medical wiki containing >1,000 pages, 250+ conditions, 5,455 treatment regimens, and 6,950 clinical studies [35]. We cleaned anti-neoplastic regimen efficacy data, linked PubMed publications and ClinicalTrials.gov data through April 2025, and publicly released the structured dataset on

[1]Given workshop limited length, we will only discuss the results of the deep research part.

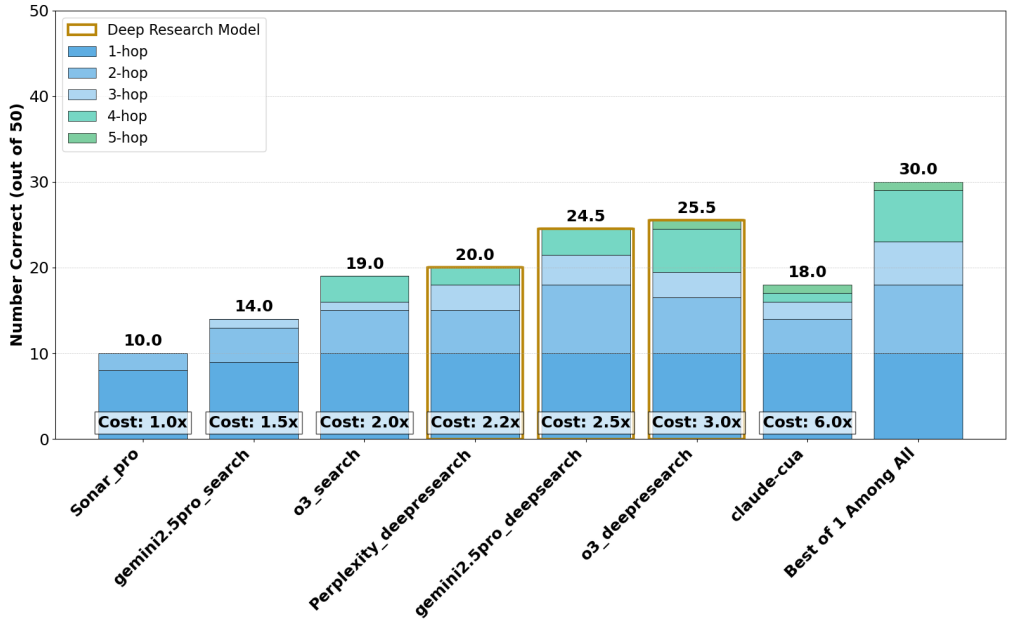

Figure 2: Overall performance of MedBrowseComp. Costs are rough estimations sorted along the x-axis. **Best of 1 Among All aggregate all measured models and their outputs.** Half point given to a specific 2-hop question where the model retrieved the sub-entities' company name instead.

HuggingFace. To create evaluation questions, we excluded multi-publication trials for disambiguation, integrated FDA Orange Book data (April 2025), and retained only trials with drugs matching standard generic expressions. Manual verification and deduplication by authors yielded 121 trials with verified metadata, efficacy data, FDA approval information, and Yahoo Finance stock data (Appendix Figure 3). This curated dataset enables assessment of autonomous computer use within 1-2 hops of HemOnc.org and deep research agents (detailed construction methodology in Appendix A6).

2.1 Model Selection and other Details

We evaluate a range of systems, from models with easy API access and systems without API access, and one Computer Use Agent system. For models with an easy API with accessible cost, we evaluated the full set, which we refer to as MedBrowseComp-605. For models without easy API acccess and/or inaccessible costs, we evaluate against a smaller set which we refer to as MedBrowseComp-50 [2]. Detailed model/system descriptions are in the appendix A.1. For answer verification, we employed an automated judge powered by GPT 4.1 mini-2025-04-14. The judging prompt was adapted from existing refined evaluation templates[6, 28]. Two annotators manually answered MedBrowseComp-50 and achieved 100% inter-annotator agreement and 98% agreement with GPT 4.1 mini.

3 Results

Figure 2 summarizes accuracy on MedBrowseComp-50. Across all systems, performance decays monotonically with hop count, corroborating prior evidence that long-horizon web navigation remains an open challenge for frontier LLM agents. Nevertheless, deep research variants–agents that allow iterative browsing steps rather than a single query—had improved performance. For example, O3 deepresearch answers 25.5/50 questions correctly, a 34% relative gain over O3 search (19/50); Gemini-2.5-pro deepsearch shows 75% improvement over its single-shot analogue (24.5 vs. 14). These gains are most pronounced on the hardest 4- and 5-hop splits, where deep research agents more than double the baseline accuracy.

[2]Authors put each of the queries into each application, got responses, and graded the final outputs. All questions in MedBrowseComp-50 cannot be answered with NA.

Consistent trends are observed in the MedBrowseComp-605 results, for which only models using parametric memories and RAG framework were evaluated. The performance of models with just parametric memories is notably poor across the majority of tasks, which aligns with our intention to create a challenging benchmark. RAG improves overall performance, but its benefit diminishes with increasing hops. On MedBrowseComp-605, we observe the same core patterns when comparing "bare" parameter-only models—i.e., those relying exclusively on their internal (parametric) memory—with retrieval-augmented variants. In isolation, parameter-only systems struggle across nearly every hop depth, confirming that our benchmark delivers the intended level of difficulty. When applying RAG, models achieve notable improvements on shallow questions (1–2 hop); for example, GPT-4.1 gains 30% and Gemini-2Flash gains 41%, while Gemini-2.5Pro shows smaller boosts of 6.7% and 7.4%. However, these gains diminish beyond the third hop, and by the 4–5 hop levels RAG offers virtually no advantage over parameter-only models. However, the utility of retrieval diminishes beyond the third hop: by the 4- and 5-hop levels, RAG provides virtually no additional benefit over the bare model. The detailed results of MedBrowseComp-50 and MedBrowseComp-605 are in the Appendix A.4.

System-wise, O3 deepresearch and Gemini-2.5-pro deepsearch constitute the frontier, trailing only the upper bound of the 'Best of 1' (30/50) that selects post hoc the single best model/system answer per question.[3] O3 deepresearch and Gemini-2.5-pro deepsearch's advantage over specialized retrieval systems such as Sonar Pro (10/50) or Perplexity deepresearch (20/50) may suggest that contemporary instruction-tuned LLMs can outperform purpose-built agentic pipelines when granted autonomous browsing. However, even the best system falls short of perfect accuracy, underscoring the need for research in planning, tool use, and hallucination suppression in complex biomedical information-seeking tasks. Appendix Table 5 shows some common error modes in examples.

4 Conclusion and Future Work

Limitations: This work has three main limitations. First, all responses were judged automatically using an LLM rubric with only light human auditing; while agreement with human verification was good, reliance on machine judgment introduces potential bias. Second, benchmark construction and experimentation required substantial compute and subscription resources ($3,690 total: $320 on Perplexity, $450 on Gemini 2.5 Pro, $2,500 on Claude Sonnet 3.7 CUA, and smaller costs for ChatGPT Pro, advanced reasoning, and GPT-4.1 mini judging). Third, real-world clinical validation was not performed. MEDBROWSECOMP covers only a small portion of the vast medical knowledge base, though such focused, verifiable benchmarks remain necessary given the lack of expert-curated ground truth across the domain.

Future Work: Several extensions are planned. Beyond single-field extraction, we aim to design tasks involving multi-paragraph justification, concordance with clinical guidelines, and trend analysis in financial and regulatory contexts, all requiring deeper reasoning. We will also evaluate tool-augmented agents, such as lightweight adapters for PDF parsing, table detection, or ClinicalTrials.gov integration, and test them in AI-IDE environments (e.g., Cursor, Windsurf) where agents must sustain state, debug, and refactor code across tasks. Finally, a key direction is studying agentic systems with human-in-the-loop workflows and benchmarking their comparative performance against clinicians in realistic decision-making settings.

Conclusion: We introduced MEDBROWSECOMP, the largest verifiable benchmark for evaluating deep research and computer use agents in the medical domain. Unlike contrived tasks, each benchmark query is grounded in a clinically meaningful information-seeking scenario. Experimental results reveal clear capability gaps: retrieval-augmented pipelines answer roughly half as many queries as deep research systems, and no evaluated approach—including computer use agents—achieves more than 40% accuracy on multi-hop questions. These findings underscore the need for more robust, tool-integrated, and clinically validated agentic systems, and position MEDBROWSECOMP as a foundation for driving progress in this area.

[3] Unlike prior work showing that repeatedly sampling from a single system can boost performance, our cross-model, test-time compute extension demonstrates even greater gains in overall accuracy [28]. However, the computational expense of querying multiple distinct agents for every question is substantial.

References

[1] Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models, 2023.

[2] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *NeurIPS*, 2021.

[3] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021.

[4] Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Zhu Lei, and Michael Lingzhi Li. Benchmarking large language models on CMExam - a comprehensive chinese medical exam dataset. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[5] Qichen Ye, Junling Liu, Dading Chong, Peilin Zhou, Yining Hua, Fenglin Liu, Meng Cao, Ziming Wang, Xuxin Cheng, Zhu Lei, and Zhenhua Guo. Qilin-med: Multi-stage knowledge injection advanced medical large language model, 2024.

[6] Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang ... Summer Yue, Alexandr Wang, and Dan Hendrycks. Humanity's last exam, 2025. URL `https://arxiv.org/abs/2501.14249`.

[7] 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.

[8] Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, et al. Freshllms: Refreshing large language models with search engine augmentation. *arXiv preprint arXiv:2310.03214*, 2023.

[9] Salaheddin Alzubi, Creston Brooks, Purva Chiniya, Edoardo Contente, Chiara von Gerlach, Lucas Irwin, Yihan Jiang, Arda Kaz, Windsor Nguyen, Sewoong Oh, et al. Open deep search: Democratizing search with open-source reasoning agents. *arXiv preprint arXiv:2503.20201*, 2025.

[10] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. Webgpt: Browser-assisted question-answering with human feedback, 2022. URL `https://arxiv.org/abs/2112.09332`.

[11] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models, 2023. URL `https://arxiv.org/abs/2305.16291`.

[12] Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. Realtime qa: What's the answer right now?, 2024. URL `https://arxiv.org/abs/2207.13332`.

[13] Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, 36:28091–28114, 2023.

[14] Yiheng Xu, Zekun Wang, Junli Wang, Dunjie Lu, Tianbao Xie, Amrita Saha, Doyen Sahoo, Tao Yu, and Caiming Xiong. Aguvis: Unified pure vision agents for autonomous gui interaction. *arXiv preprint arXiv:2412.04454*, 2024.

[15] Vardaan Pahuja, Yadong Lu, Corby Rosset, Boyu Gou, Arindam Mitra, Spencer Whitehead, Yu Su, and Ahmed Awadallah. Explorer: Scaling exploration-driven web trajectory synthesis for multimodal web agents. *arXiv preprint arXiv:2502.11357*, 2025.

[16] M. Wornow, Y. Xu, R. Thapa, et al. The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine*, 6:135, 2023. doi: 10.1038/s41746-023-00879-8.

[17] Shan Chen, Marco Guevara, Shalini Moningi, Frank Hoebers, Hesham Elhalawani, Benjamin H. Kann, Fallon E. Chipidza, Jonathan Leeman, Hugo J. W. L. Aerts, Timothy Miller, Guergana K. Savova, Raymond H. Mak, Maryam Lustberg, Majid Afshar, and Danielle S. Bitterman. The impact of responding to patient messages with large language model assistance, 2023.

[18] Shan Chen, Benjamin H. Kann, Michael B. Foote, Hugo J. W. L. Aerts, Guergana K. Savova, Raymond H. Mak, and Danielle S. Bitterman. Use of artificial intelligence chatbots for cancer treatment information. *JAMA Oncology*, 9(10):1459–1462, 2023. doi: 10.1001/jamaoncol.2023.2954. URL `https://jamanetwork.com/journals/jamaoncology/article-abstract/2804562`.

[19] Hongjian Zhou, Fenglin Liu, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S. Chen, Peilin Zhou, Junling Liu, Yining Hua, Chengfeng Mao, Chenyu You, Xian Wu, Yefeng Zheng, Lei Clifton, Zheng Li, Jiebo Luo, and David A. Clifton. A survey of large language models in medicine: Progress, application, and challenge, 2024.

[20] Kun-Hsing Yu, Elizabeth Healey, Tze-Yun Leong, Isaac S Kohane, and Arjun K Manrai. Medical artificial intelligence and human values. *New England Journal of Medicine*, 390(20): 1895–1904, 2024.

[21] Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Yong Cheng, Le Hou, Albert Webson, Kavita Kulkarni, S Sara Mahdavi, Christopher Semturs, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S Corrado, Yossi Matias, Alan Karthikesalingam, and Vivek Natarajan. Towards conversational diagnostic ai, 2024. URL `https://arxiv.org/abs/2401.05654`.

[22] Jack Gallifant, Majid Afshar, Saleem Ameen, Yindalon Aphinyanaphongs, Shan Chen, Giovanni Cacciamani, Dina Demner-Fushman, Dmitriy Dligach, Roxana Daneshjou, Chrystinne Fernandes, et al. The tripod-llm reporting guideline for studies using large language models. *Nature Medicine*, pages 1–10, 2025.

[23] Yubin Kim, Hyewon Jeong, Shan Chen, Shuyue Stella Li, Mingyu Lu, Kumail Alhamoud, Jimin Mun, Cristina Grau, Minseok Jung, Rodrigo Gameiro, et al. Medical hallucinations in foundation models and their impact on healthcare. *arXiv preprint arXiv:2503.05777*, 2025.

[24] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams, 2020. URL `https://arxiv.org/abs/2009.13081`.

[25] João Matos, Shan Chen, Siena Placino, Yingya Li, Juan Carlos Climent Pardo, Daphna Idan, Takeshi Tohyama, David Restrepo, Luis F. Nakayama, Jose M. M. Pascual-Leone, Guergana Savova, Hugo Aerts, Leo A. Celi, A. Ian Wong, Danielle S. Bitterman, and Jack Gallifant. Worldmedqa-v: a multilingual, multimodal medical examination dataset for multimodal language models evaluation, 2024. URL `https://arxiv.org/abs/2410.12722`.

[26] Jiwen Zhang, Jihao Wu, Yihua Teng, Minghui Liao, Nuo Xu, Xiao Xiao, Zhongyu Wei, and Duyu Tang. Android in the zoo: Chain-of-action-thought for gui agents. *arXiv preprint arXiv:2403.02713*, 2024.

[27] Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh J Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, et al. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems*, 37:52040–52094, 2024.

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

[29] Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: a benchmark for general ai assistants. In *The Twelfth International Conference on Learning Representations*, 2023.

[30] Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Linhai Zhang, Yulan He, Deyu Zhou, Pengjun Xie, et al. Webwalker: Benchmarking llms in web traversal. *arXiv preprint arXiv:2501.07572*, 2025.

[31] Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. *arXiv preprint arXiv:2409.12941*, 2024.

[32] Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*, 2024.

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

[34] Yixiao Song, Katherine Thai, Chau Minh Pham, Yapei Chang, Mazin Nadaf, and Mohit Iyyer. Bearcubs: A benchmark for computer-using web agents. *arXiv preprint arXiv:2503.07919*, 2025.

[35] Jeremy L. Warner, Dmitry Dymshyts, Christian G. Reich, Michael J. Gurley, Harry Hochheiser, Zachary H. Moldwin, Rimma Belenkaya, Andrew E. Williams, and Peter C. Yang. Hemonc: A new standard vocabulary for chemotherapy regimen representation in the omop common data model. *Journal of Biomedical Informatics*, 96:103239, Aug 2019. doi: 10.1016/j.jbi.2019.103239. URL `https://pmc.ncbi.nlm.nih.gov/articles/PMC6697579/`.

Ethics MEDBROWSECOMP is built solely from publicly available, non-identifiable sources, so no protected health information is exposed. Its results are intended for research benchmarking and must not be interpreted as clinical performance guarantees; any real-world deployment requires qualified human oversight. We acknowledge potential corpus-level biases and plan bias audits and broader source diversification in future releases.

A Appendix

A.1 Detailed Benchmark Setting

Model	Mode	Test Time	Version	Verification Method
		MedBrowseComp 50		
O3	search	April 29th, 2025	Pro subscription	llm + human
O3	deepresearch	April 29 – May 1st, 2025	Pro subscription	llm + human
Gemini2.5pro	search	May 1st, 2025	api - 03/25	llm + human
Gemini2.5pro	deepresearch	April 30 – May 1st, 2025	One subscription	llm + human
Sonar pro	search	April 28th, 2025	api	llm + human
Perplexity	deepresearch	April 28th, 2025	api	llm + human
Sonnet 3.7	CUA	May 6th, 2025	api - vertex 20250219	llm + human
		MedBrowseComp 605		
Gemini2.5pro	search	May 2nd, 2025	api - 03/25	llm
GPT-4.1	search-high	May 2nd, 2025	api - gpt-4.1-mini-2025-04-14	llm
Gemini2flash	search	May 2nd, 2025	api	llm
Gemini2.5pro	search	May 1st, 2025	api - 03/25	llm
Sonar pro	search	May 2nd, 2025	api	llm
		MedBrowseComp CUA		
Sonnet 3.7	CUA	May 6th, 2025	api - vertex 20250219	llm

Table 1: Overview of test runs for MedBrowseComp benchmarks.

A.2 LLM as judge Prompts

We mostly use the same additional instruction as Humanity's Last Exam [6]. However we did some modification due to our need. First we used OpenAI structure output pipeline accompanied with a system prompt of *You are an impartial judge evaluating an AI response based on provided criteria. Respond ONLY with a valid JSON object matching the requested structure.*

We also did not include the confidence estimation as we do not have API access to many of the models.

Pydantic Schema for Structured Output

```
class JudgeOutput(BaseModel):
    extracted_final_answer: str = Field(
        description="The final exact answer extracted from the [
    response]. Put 'None' if no exact answer found."
    )
    reasoning: str = Field(
        description="Explanation of correctness based ONLY on
    comparing extracted_final_answer and correct_answer."
    )
    correct: str = Field(
        description="Must be 'yes' or 'no'."
    )
```

Grading Prompt

```
JUDGE_PROMPT = """Judge whether the following [response] to [question]
is correct or not based on the
precise and unambiguous [correct_answer] below.
[question]:  {question}
[response]:  {response}
Your judgement must be in the format and criteria specified below:
extracted_final_answer:  The final exact answer extracted from the
[response].  Put the extracted answer
as 'None' if there is no exact, final answer to extract from the
response.
[correct_answer]:  {correct_answer}
reasoning:  Explain why the extracted_final_answer is correct or
incorrect based on [correct_answer],
focusing only on if there are meaningful differences between
[correct_answer] and the
extracted_final_answer.  Do not comment on any background to the
problem, do not attempt to solve
the problem, do not argue for any answer different than
[correct_answer], focus only on whether the
answers match.
correct:  Answer 'yes' if extracted_final_answer matches the
[correct_answer] given above, or is within
a small margin of error for numerical problems.  Answer 'no' otherwise,
i.e.  if there is
any inconsistency, ambiguity, non-equivalency, or if the extracted
answer is incorrect."""
```

A.3 Computer Use Agent Prompt

The prompt remains largely identical to the original version provided by Anthropic, with a single minor addition highlighted in gray.

This addition is a concise instruction reinforcing that the agent should act independently and refrain from requesting clarification or human assistance during execution, promoting model autonomy.

<SYSTEM_CAPABILITY>

- You are utilising an Ubuntu virtual machine using {platform.machine()} architecture with internet access.
- You can feel free to install Ubuntu applications with your bash tool. Use curl instead of wget.
- To open Firefox, please just click on the Firefox icon. Note, firefox-esr is what is installed on your system.
- Using bash tool you can start GUI applications, but you need to set export DISPLAY=:1 and use a subshell. For example, (DISPLAY=:1 xterm &). GUI apps run with bash tool will appear within your desktop environment, but they may take some time to appear. Take a screenshot to confirm it did.
- When using your bash tool with commands that are expected to output very large quantities of text, redirect into a tmp file and use str_replace_editor or grep -n -B <lines before> -A <lines after> <query> <filename> to confirm output.
- When viewing a page it can be helpful to zoom out so that you can see everything on the page. Either that, or make sure you scroll down to see everything before deciding something isn't available.
- When using your computer function calls, they take a while to run and send back to you. Where possible/feasible, try to chain multiple of these calls all into one function calls request.
- The current date is {datetime.today().strftime('%A, %B %-d, %Y')}.

</SYSTEM_CAPABILITY>

<IMPORTANT>

- Never ask the user for help or to clarify. You are the assistant and you should be able to figure out what to do.
- When using Firefox, if a startup wizard appears, IGNORE IT. Do not even click “skip this step.” Instead, click on the address bar where it says “Search or enter address,” and enter the appropriate search term or URL there.
- If the item you are looking at is a PDF, and after taking a single screenshot of the PDF it seems that you want to read the entire document instead of trying to continue to read the PDF from your screenshots + navigation, determine the URL, use curl to download the PDF, install and use pdftotext to convert it to a text file, and then read that text file directly with your StrReplaceEditTool.

</IMPORTANT>

A.4 Deep Research Agent Results

Table 2: Detailed Performance of Frontier Systems on MedBrowseComp 50 - For this subset, we selected where the questions cannot be answered by NA

Question Depth	O3		Gemini2.5pro		Perplexity		Claude-CUA
	search	deep	search	deep	search	deep	
1-hop (n=10)	10/10 (**100.0%**)	10/10 (**100.0%**)	9/10 (90.0%)	10/10 (**100.0%**)	8/10 (80.0%)	10/10 (**100.0%**)	10/10 (**100.0%**)
2-hop (n=10)	5/10 (50.0%)	6.5/10 (65.0%)	4/10 (40.0%)	**8**/10 (**80.0%**)	2/10 (20.0%)	5/10 (50.0%)	4/10 (40.0%)
3-hop (n=10)	1/10 (10.0%)	3/10 (30.0%)	1/10 (10.0%)	**3.5**/10 (**35.0%**)	0/10 (0.0%)	3/10 (30.0%)	2/10 (20.0%)
4-hop (n=10)	3/10 (30.0%)	**5**/10 (**50.0%**)	0/10 (0.0%)	3/10 (30.0%)	0/10 (0.0%)	2/10 (20.0%)	1/10 (10.0%)
5-hop (n=10)	0/10 (0.0%)	**1**/10 (**10.0%**)	0/10 (0.0%)	0/10 (0.0%)	0/10 (0.0%)	0/10 (0.0%)	1/10 (10.0%)
Total (n=50)	19/50 (38.0%)	**25.5**/50 (**51.0%**)	14/50 (28.0%)	24.5/50 (49.0%)	10/50 (20.0%)	20/50 (40.0%)	18/50 (36.0%)

The benchmark we've created is challenging, as demonstrated by results on MedBrowseComp 50 and more detailed results on MedBrowseComp 605 on the following page. It's designed to push models beyond just recognizing patterns or guessing from context. Instead, it asks them to follow a chain of reasoning across multiple steps (or "hops") through a medical knowledge base. When you strip away the ability to say "Not applicable" or avoid answering (referred to here as REAL accuracy), the results show just how hard this is. Even the strongest model, Gemini 2.5 Pro (search) , only gets about 24.5% of the answers right under these strict conditions. That might not sound like much, but it still makes it the clear leader in this group.

What is especially telling is how performance drops off with each additional hop. For example, Gemini 2.5 Pro does well on 1-hop questions (76%), where the answer is often directly stated. But by the time you get to 4-hop or 5-hop questions — where the model has to link together several pieces of information in sequence — even this model struggles. On REAL accuracy for 4-hop questions, Gemini 2.5 Pro only gets 5.1% , and for 5-hop, it's essentially zero. This shows that while models may look good on simple tasks, chaining together multiple steps of reasoning is still a major challenge.

In short, we hope this benchmark doesn't let models take shortcuts. We want to force them to dig into real medical knowledge and reason carefully. And based on these results, there's still a long way to go before we can fully trust AI systems to handle complex, multi-step medical reasoning without supervision.

Table 3: Detailed Performance of Models on MedBrowseComp 605 | Note that SonarPro-param is blank here due to the lack of non-search options from perplexity.

Question Depth	GPT-4.1		SonarPro		Gemini2Flash		GeminiPro	
	param	search	param	search	param	search	param	search
1-hop (n=121)	24/121 (19.8%)	19/121 (15.7%)	–	63/121 (52.1%)	26/121 (21.5%)	67/121 (55.4%)	10/121 (8.3%)	**92**/121 (**76.0%**)
2-hop (n=121)	5/121 (4.1%)	5/121 (4.1%)	–	8/121 (6.6%)	2/121 (1.7%)	7/121 (5.8%)	4/121 (3.3%)	**13**/121 (**10.7%**)
3-hop (n=121)	1/121 (0.8%)	1/121 (0.8%)	–	2/121 (1.7%)	2/121 (1.7%)	1/121 (0.8%)	9/121 (7.4%)	**4**/121 (**3.3%**)
4-hop (n=121)	60/121 (49.6%)	42/121 (34.7%)	–	70/121 (57.9%)	60/121 (49.6%)	48/121 (39.7%)	39/121 (32.2%)	**49**/121 (**40.5%**)
5-hop (n=121)	15/121 (12.4%)	12/121 (9.9%)	–	15/121 (12.4%)	23/121 (19.0%)	15/121 (12.4%)	18/121 (14.9%)	**24**/121 (**19.8%**)
Total (n=605)	105/605 (17.3%)	80/605 (13.2%)	–	158/605 (26.1%)	113/605 (18.7%)	138/605 (22.8%)	80/605 (13.2%)	**182**/605 (**30.1%**)

Table 4: REAL Accuracy for MedBrowseComp605 (Excluding NA-like Correct Answers): Models are evaluated only on questions where the correct answer is applicable. NA-like responses (e.g., "Not applicable") are excluded from scoring.

Question Depth	GPT-4.1		SonarPro		Gemini2Flash		GeminiPro	
	param	search	param	search	param	search	param	search
1-hop (n=121)	24/121 (19.8%)	19/121 (15.7%)	–	63/121 (52.1%)	26/121 (21.5%)	67/121 (55.4%)	10/121 (8.3%)	**92**/121 (**76.0%**)
2-hop (n=121)	5/121 (4.1%)	5/121 (4.1%)	–	8/121 (6.6%)	2/121 (1.7%)	7/121 (5.8%)	4/121 (3.3%)	**13**/121 (**10.7%**)
3-hop (n=121)	1/121 (0.8%)	1/121 (0.8%)	–	2/121 (1.7%)	2/121 (1.7%)	1/121 (0.8%)	9/121 (7.4%)	**4**/121 (**3.3%**)
4-hop (n=39)	0/39 (0.0%)	0/39 (0.0%)	–	0/39 (0.0%)	0/39 (0.0%)	0/39 (0.0%)	0/39 (0.0%)	**2**/39 (**5.1%**)
5-hop (n=51)	0/51 (0.0%)	0/51 (0.0%)	–	1/51 (2.0%)	1/51 (2.0%)	0/51 (0.0%)	0/51 (0.0%)	**0**/51 (**0.0%**)
Total (n=453)	30/453 (6.6%)	25/453 (5.5%)	–	74/453 (16.3%)	31/453 (6.8%)	75/453 (16.6%)	23/453 (5.1%)	**111**/453 (**24.5%**)

A.5 Deep Research Agents Common Error Mode

Table 5: Examples of common errors among deep research systems on MedBrowseComp-50.

Error type & question (paraphrased)	Explanation
Type: Inefficient tool allocation **Question:** "For clinical trial `NCT00974311`, review the more effective regimen ingredients and identify which ingredient starts with the letter 'E'. Then, determine which pharmaceutical company received the most recent FDA approval (until December 2024) for this identified ingredient. If this company is listed on any US stock market, provide the stock ticker symbol and opening stock price on the FDA approval date."	Agent exhausted its tool-call quota on preliminary tasks (trial verification, news validation) instead of reserving sufficient calls for retrieving critical financial data (stock price lookup). Consequently, the final essential query regarding the stock price was left unanswered, with the agent stating financial data was unavailable in the provided materials.
Type: Poor source selection **Question:** "In trial `NCT00974311`, identify the effective ingredient beginning with 'E' and return its FDA exclusivity date (overall approval only, in `MM-DD-YYYY` format)."	The agent cited secondary press releases and FDA news items (e.g., Pfizer announcements) instead of querying the FDA *Orange Book*, the authoritative source for exclusivity information. As a result, it either reported the initial approval date rather than the exclusivity expiry or claimed the exclusivity data were unavailable.
Type: Unable to parse long context tables **Question:** "For clinical trial `NCT00720512`, among the more effective regimen ingredients, identify which ingredient starts with the letter 'I'. Then, for this ingredient last approved up to December 2024, provide its patent expiration date (overall FDA approval only). Return only `YYYY`."	The agent attempted to extract patent-expiry data from a multi-page Orange Book PDF containing dense tables. Because it did not robustly parse the table structure, it surfaced only partial approval milestones (e.g., accelerated/full approvals for irinotecan) and never captured the patent-expiration year requested, leading to an incomplete answer.

A.6 Details on Benchmark Curation

To construct the MedBrowseComp dataset, we leverage the comprehensive hematology and oncology database available from HemOnc.org and work with its editors. HemOnc.org is the largest freely available medical wiki in the field of hematology/oncology, established to address the challenge oncologists routinely face navigating complex treatment regimens and rapidly evolving standards of care. This comprehensive resource covers over 1,000 pages of specialized content, including more than 250 hematologic and oncologic conditions, 5,455 detailed treatment regimens, and 6,950 referenced clinical studies, all curated by physicians with verifications. The platform catalogs approved systemic antineoplastic therapy agents, supportive medications, standard-of-care regimens, and references to primary literature, organized within a standardized ontology framework available through the HemOnc Dataverse [35]. Our first step involved cleaning anti-neoplastic regimen efficacy data, linking each case with corresponding PubMed publications, and associated clinical trial information sourced from ClinicalTrials.gov, with data collected up to April 2025. The fully cleaned and structured version of this dataset has been publicly released on HuggingFace to facilitate broader community engagement, further development, and external validation.

To create our specific evaluation questions, we narrowed our dataset by excluding trials linked to multiple PubMed publications to maintain clarity and verifiability. Subsequently, we integrated regimen-specific drug information with FDA Orange Book data as of April 2025. To maintain data consistency, only trials with regimens containing drugs easily matched through standard generic regular expressions were included. A manual verification and deduplication process, led by author SC, was conducted to ensure accuracy and reduce redundancy, culminating in a refined set of 121 trials. Each of these trials has clearly defined trial metadata, verified regimen efficacy data, detailed

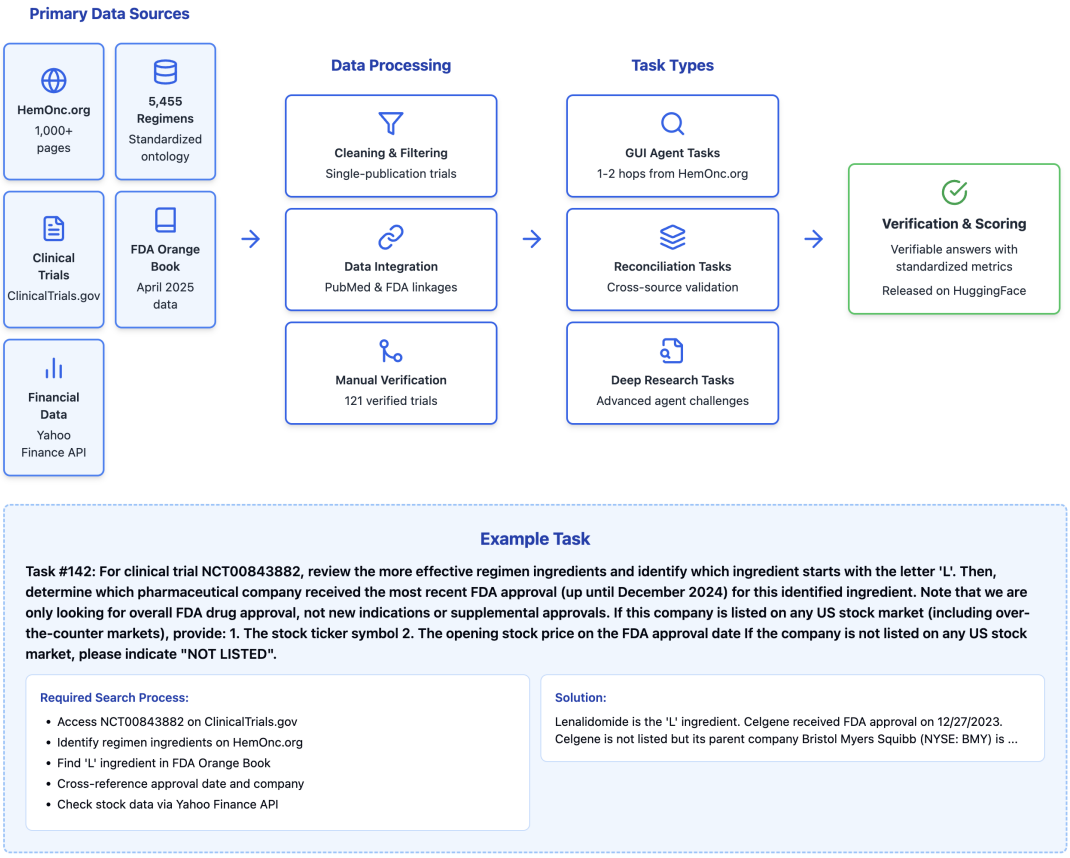

Figure 3: Overall workflow of the curation of MedBrowseComp.

FDA drug approval information, and the corresponding financial market data obtained from Yahoo Finance API for associated stock pricing, as Appendix Figure 3 shows.

From this carefully curated dataset, we developed our benchmark designed explicitly to assess (1) autonomous CUA within one to two hops of HemOnc.org's webpage, and deep research agents.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Contributions are clearly enumerated at the end of the introduction, highlighting results and resources that can be found within the manuscript.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: A dedicated limitations section can be found at the end of the paper.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - A separate "Limitations" section in the paper clearly enumerates the key limitations of this paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA] .

Justification: No theoretical results are presented in this piece. Any calculations have associated equations in-line and are referenced as such.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: All code is available in a public repository.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: A detailed README has been provided within each repository folder describing the steps required to reproduce or extend the current work. All final counts, outputs, and LLM as judge results are available for download on the public website.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer:[Yes]

Justification: While no training or tuning was conducted, we provided all our code, settings and outputs.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: Given the nature of benchmarking and excessive costs plus we do not have API access to many of the services. All of the results we provided are pass@1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: They are all included in our conclusion section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The authors of this study have read, and confirm this study conforms with every aspect of the Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We do not think our paper holds many negative societal impacts. We did include an ethic section to discuss in Section 5. And, we are eager to discuss and include if proper during the rebuttal.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: All models and datasets utilized in this study are already publicly available. However, to prevent pre-training contamination in from scraping GitHub, we include an encoded version of our dataset publicly.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets are open access and comply with the copyright and terms of service under Apache 2.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Details of the datasets, code, and findings are all available on our website. We have also provided a blog on this website with a more user-friendly explanation of the approach and findings. this aims to increase accessibility of the results to a broader audience.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: While no crowdsourcing was utilized, details of how our results are gathered and validated by the research team are provided.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLM is not used as the core methodology here.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.