MedBrowseComp: Benchmarking Medical Deep Research and Computer Use

Shan Chen[1,2,3]*, **Pedro Moreira**[4,5]*, **Yuxin Xiao**[4]
Sam Schmidgall[6], **Jeremy Warner**[7,8], **Hugo Aerts**[1,2,9]
Thomas Hartvigsen[10], **Jack Gallifant**[1,2], **Danielle S. Bitterman**[1,2,3,8]†

[1]Harvard, [2]Mass General Brigham, [3]Boston Children's Hospital, [4]MIT, [5]Universitat Pompeu Fabra, [6]Johns Hopkins University, [7]Brown University, [8]HemOnc.org, [9]Maastricht University, [10]University of Virginia

https://huggingface.co/datasets/AIM-Harvard/MedBrowseComp
https://github.com/shan23chen/MedBrowseComp

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 live, domain-specific knowledge bases. MedBrowseComp holds 1,000+ human-curated questions that mirror clinical scenarios in which practitioners must reconcile fragmented or conflicting information to reach an up-to-date conclusion. Applying MedBrowseComp to frontier agentic systems reveals **marked performance shortfalls as low as 10%**. These findings expose a critical gap between current LLM capabilities and the rigor demanded in clinical settings. MedBrowseComp exposes the strengths and weaknesses of current agentic systems, offering a testbed for reliable medical information seeking and clear goals for future model and toolchain upgrades.

The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance

1 Introduction

LLMs have now effectively saturated most static, knowledge-based benchmarks, diminishing the value of these tasks to provide new insights and push the field forward[1–6]. However, this evolution exposes an *evaluation gap*: legacy leaderboards track static knowledge recall, whereas agentic systems should *plan*, *browse*, and *synthesize* fresh evidence in real time. To move beyond this plateau, the community is pivoting toward *agentic* systems that actively browse the web, retrieve real-time evidence, and reason over information outside their frozen parameter memories [7–9]. The progression from chatbots to reasoners and ultimately to autonomous agents promises to enable access to real-time data, to allow models to tackle questions outside their pretraining knowledge, and perform complex information gathering tasks previously exclusive to humans [10–12].

*Co-first authors: Shan Chen and Pedro Moreira
†Corresponding author: dbitterman@bwh.harvard.edu

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

The potential impact of web-enabled agents is enormous. In principle, a sufficiently-capable AI agent should be able to retrieve any well-specified fact from the open web, even if doing so requires navigating thousands of pages [13–15]. This promise is especially compelling in medicine, where research, clinical decision support, and patient education all depend on the integration of the most current, specific, and accurate information from heterogeneous sources such as journal articles, clinical trial registries, treatment guidelines, and drug databases [16–22]. Yet despite rapid progress, the community lacks a unified benchmark for systematically evaluating whether agents can perform such complex, multi-source medical retrieval at scale.

As LLMs become the backbone of autonomous agentic systems, rigorous, domain-specific evaluation is critical. Contemporary agent models frequently "hallucinate," generating confident but unsubstantiated or factually incorrect statements [23]. In high-stakes fields like medicine, these errors can misinform clinicians or patients and erode trust. Therefore, benchmarks must test not only a model's reasoning and navigation skills but also its ability to ground each answer in verifiable evidence [18, 19, 23]. A robust framework that measures evidence-based accuracy, navigation efficiency, and citation fidelity would directly quantify how well agentic systems incorporate best available information, closing the gap between impressive demos and safe, real-world deployment.

Evaluating an agent's deep research and computer use competence is fundamentally different from testing its ability to recall static facts. Popular medical benchmarks such as MMLU, MedQA, and WorldMedQA test information that can be memorized during pre-training or retrieved from a single authoritative page. Frontier models now achieve near-ceiling scores on these tasks, so the benchmarks can no longer distinguish state-of-the-art systems or quantify new progress [3, 24, 25]. They are usually executed in closed or synthetic environments that sidestep the real-world hurdles of live web interaction—pagination, obsolete links, contradictory evidence, and shifting page layouts [26, 27].

In medicine, clinicians and researchers must integrate the latest study results, guideline updates, and safety advisories scattered across heterogeneous sites. To understand and compare agents' abilities to augment such tasks, new deep research and computer use benchmarks reflective of real-world, up-to-date tasks are urgently needed. A benchmark that forces agents to conduct multi-hop, evidence-grounded searches on the open Web should therefore serve two complementary purposes: **1)** Directly measure whether agents can navigate, filter, and reconcile real-world information. **2)** Dynamically stress-test systems as underlying evidence evolved, long after static benchmarks saturate.

To address this gap, we introduce MedBrowseComp, a new benchmark specifically designed to evaluate the capabilities of AI agents performing complex information retrieval tasks within the medical domain via web browsing. MedBrowseComp measures an agent's ability to accurately navigate the web—including general web pages, specialized medical websites, databases, and potentially document formats like wikis, PDFs—to locate verifiable medical facts.

MedBrowseComp's design is inspired by the BrowseComp benchmark [28]; it focuses on fact-seeking questions where the answers are short, objective, and easily verifiable, simplifying the evaluation process and enhancing its reliability. **We designed this challenging benchmark collaboratively with physicians using HemOnc.org, one of the largest structured wiki information resources maintained weekly by oncologists for the past 6 years.** The benchmark is designed to enable automated, dynamic updating as information resources evolve. State-of-the-art Deep Research systems and Computer Use Agents(CUA) in May 2025 achieve less than 50% accuracy overall on MedBrowseComp, with less than 10% in the two hardest sets of questions.

The primary contributions of this work are:

The MedBrowseComp Dataset: A novel, curated collection of challenging medical fact-seeking questions, each requiring web browsing and resulting in a short, verifiable answer. The pioneer in curating a comprehensive benchmark that utilizes linked domain knowledge.

Baseline Performance Analysis: An empirical evaluation of various state-of-the-art LLMs and agentic systems on MedBrowseComp, providing initial benchmarks and highlighting the specific difficulties encountered in medical information retrieval.

Demonstration of Capability Gaps: Evidence showing the gap between the capabilities of general-purpose browsing agents and specialized skills on complex medical information-seeking tasks.

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.

2 Related Work

One of the first comprehensive efforts to advance AI evaluation for general-purpose tool use and web browsing is GAIA [29]. GAIA is a carefully curated testbed for "General AI Assistants," combining tasks that require multi-modal input, open-ended reasoning, and the ability to query external tools. It is simple for humans, but it was hard for AI systems at the time. Subsequent work began to highlight the importance of structured, multi-step web traversal. WebWalker introduced a dual-agent framework that separately handles horizontal browsing across different webpages and deep vertical navigation through website hierarchies [30, 27]. Its companion benchmark, WebWalkerQA, comprises realistic multi-hop questions from domains like education and organizational websites, emphasizing the challenge of distributing information across intricate hyperlink structures. Around the same time, FRAMES took a different angle by systematically evaluating Retrieval-Augmented Generation (RAG) systems for factual correctness, retrieval quality, and reasoning [31]. While WebWalker and FRAMES address structured exploration and pipeline analysis, SimpleQA focuses more narrowly on short-answer factual correctness, specifically targeting LLM hallucinations.[32] SimpleQA's adversarial question collection and stringent answer verification make it an effective tool for measuring whether models can reliably produce grounded, non-invented responses. Despite its value in exposing factual flaws, SimpleQA's tasks remain relatively easy to solve with basic web searches, limiting its utility for deeper or more specialized queries [9].

To test expert-level knowledge well beyond the range of average tasks, Humanity's Last Exam (HLE) provides a set of deeply specialized questions [6]. HLE intentionally targets areas where frontier models are known to have knowledge gaps, thereby illuminating the upper limits of AI comprehension across diverse academic disciplines. However, while HLE excels at assessing conceptual difficulty and depth of reasoning, it focuses less on web and knowledge-source navigation capabilities.

BrowseComp and its offshoot BrowseComp-ZH push AI agents to perform complex web searches for difficult-to-find facts [28, 33]. BrowseComp's 1,255 English questions use a reverse-engineered approach to ensure they are not trivially searchable: each was devised and combined from existing short answers. The leading OpenAI models have an overall accuracy lower than 10%, however AI systems with search functionalities perform better. BrowseComp-ZH extends this framework to the Chinese web, accounting for different linguistic structures, censorship constraints, and local data sources. Leading models still achieve below 10–20% on BrowseComp-ZH tasks, suggesting that

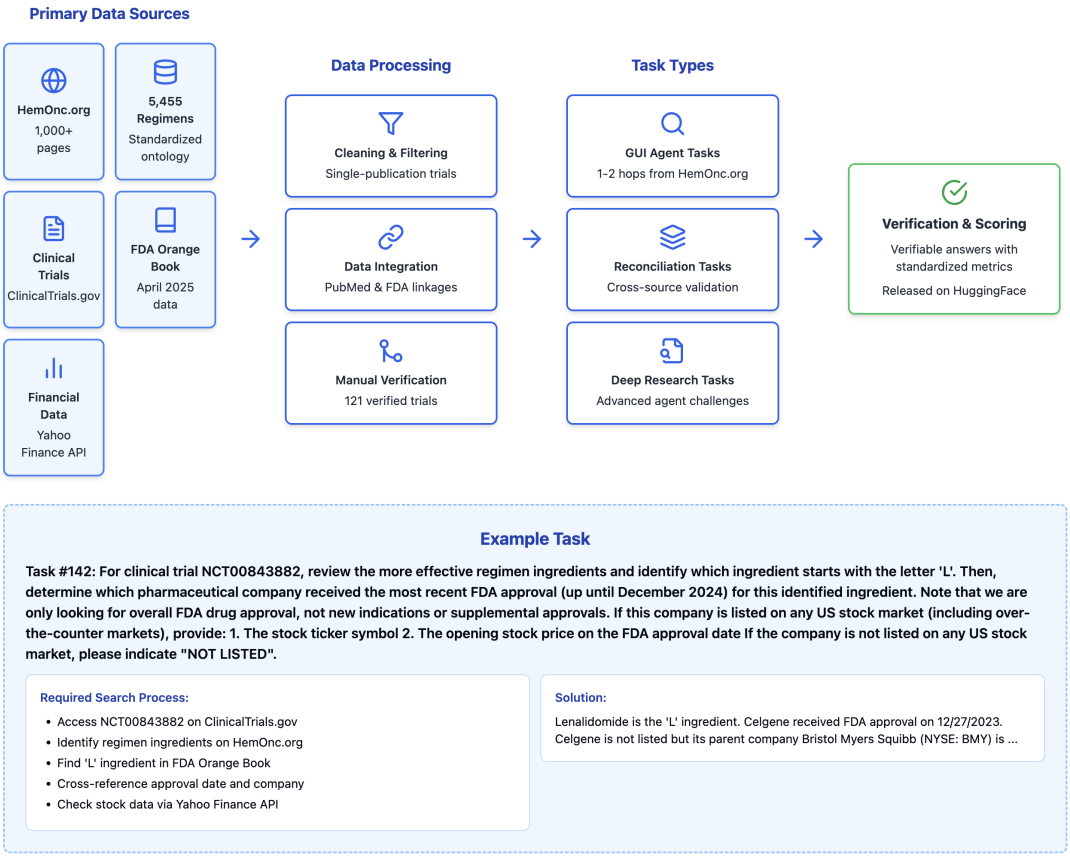

Figure 2: Overall workflow of the curation of MedBrowseComp.

domain- or region-specific peculiarities are another consideration of the difficulty of high-stakes information retrieval.

Taken together, these benchmarks outline a rapidly evolving space in AI applications, where persistent web navigation, accurate retrieval, and robust reasoning remain underdeveloped in even the most capable models [34]. MedBrowseComp bridges the gap by focusing on the specialized domain of medicine. Its 1000+ questions[3] are designed to be both practically useful and challenging, demanding persistent exploration of reputable medical sources, correct interpretation of terminology, and alignment with evidence-based standards. This unique focus is not only inherently high-stakes but also exposes the limitations of generic benchmarks, which rarely require the same level of detailed domain reasoning or cross-referencing. Moreover, MedBrowseComp's design is expandable, allowing for continual updates that track evolving medical knowledge and new online resources.

3 Methods

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

[3] 121*5 -> 605 deep research questions across five hops, 121*4 -> 484 computer use questions over four hops

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 [4]. To avoid potential dataset contamination, we encoded our final test sets with shifts and byte-wise encoding, which you can decode using the script we provide on GitHub.

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 FDA drug approval information, and the corresponding financial market data obtained from Yahoo Finance API for associated stock pricing, as Figure 2 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.

3.1 Model Selection and other Details

System Selection Rationale for Deep Research Agent 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 [5]. 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 (SC and JG), manually answered MedBrowseComp-50 and achieved 100% inter-annotator agreements and 98% agreement with GPT 4.1 mini.

Model Selection Rationale for Computer Use Agent We select Anthropic's Computer Use (using Claude 3.7 Sonnet) for evaluating browsing capabilities because, among the major commercially available GUI agents at the time, it was the only one that simultaneously offer a documented, programmable API and sandboxed environment that could be reengineered to run large batches without manual oversight, enabling us to run our experiments at scale.[6] Other UI agents—such as Bytedance's UI Tars, OpenAI's Operator, and Google's Project Mariner-were considered, but at the time of experimentation they either lacked comparable logging/automation features or required additional deployment effort that was beyond our project timeline.

Evaluation Architecture CUA is still experimental: every UI action spawns a tool call and screenshot, adding range from 2–10k tokens and extra latency. As recent benchmarks and surveys note, this overhead quickly pushes long tasks toward the model's context limit and amplifies selector failures, making deep navigation chains unreliable. Accordingly, we restrict our benchmark for UI agents to four tasks - trial ID, second author, PMID, and start date — that can be retrieved in one or two hops from HemOnc.org, providing a challenging yet attainable benchmark for GUI agents.

We implement a distributed evaluation infrastructure by adapting Anthropic's CUA codebase, transforming it from an interactive web application into a scalable, fault-tolerant parallel processing system that can be easily evaluated on personal laptops.

The evaluation back-end is organized as three cooperating services: **(1) Prompt processor** - a lightweight Python microservice that streams CSV encoded tasks into Claude 3.7 Sonnet, applying

[4] https://huggingface.co/datasets/ml4h-2025-submission/medbrowsecomp

[5] Author SC and JG put each of the queries into each application, copied out the responses, and graded the final outputs.

[6] Details on Claude's sandbox is provide in the following paragraph and the code is open sourced on our Github repository.

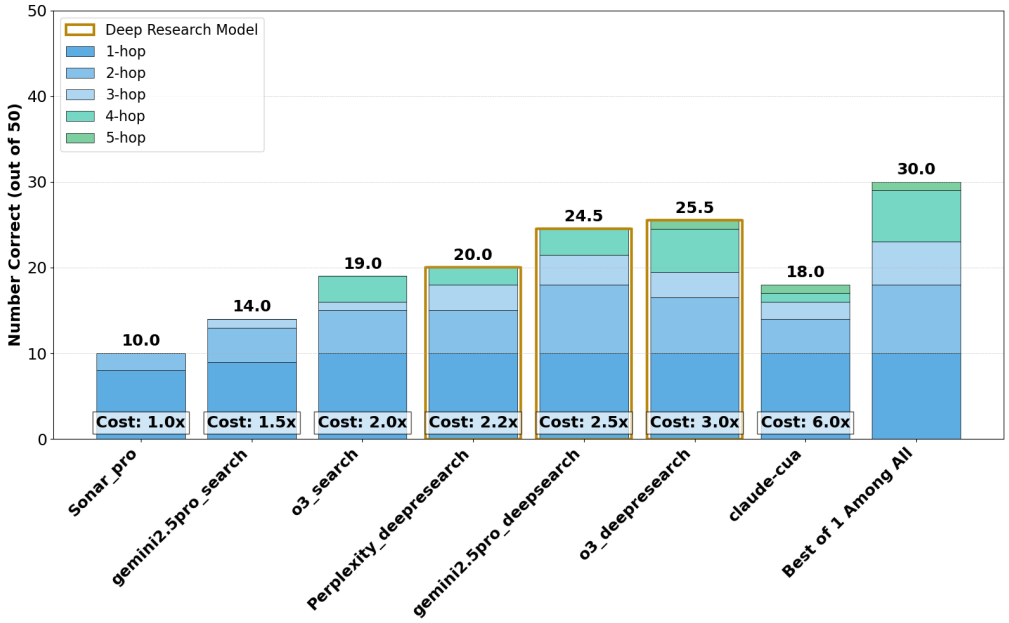

Figure 3: Overall performance of MedBrowseComp. Costs are rough estimations sorted along the x-axis. **Best of 1 Among All aggregate all measured models and their outputs.** Half point given to a specific 2-hop question where the model retrieved the sub-entities' company name instead.

up to five exponential backoff retries for transient errors (HTTP 429, timeouts, 5xx). **(2) Container orchestrator** - a Docker / Compose layer that launches N identical Ubuntu22.04 containers, mounts a shared volume, and shards the prompt file so that each container runs an independent batch without crosstalk. We ran our container in a batch of 8 on author PM's MacBook throughout the experiment. **(3) Sampling loop** - A sampling loop that calls each turn is capped at 4096 output tokens while the screenshots that are sent as input are only the three most recent, ensuring that the running context never exceeds 120k tokens - well under Sonnet's 200k window. Note that there is no limit on the number of actions the model can perform or time constraints. We retained Anthropic's default Ubuntu sandbox exactly as shipped, so the model runs in the same environment used for training — maintaining the best navigation accuracy, reproducibility, and runtime stability. The only change was a minor adjustment to the system-prompt that nudges the agent to act autonomously without asking a human for grounding or clarifications during execution which you can refer at Appendix A.3.

Lastly, we verified its difficulty by running Gemini-2.0-Flash-001 zero-shot across five runs, where it achieved under 7 % accuracy consistently on all sub-partitions of our benchmark questions.

4 Results

4.1 Deep Research Agent Results

Figure 3 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.

Consistent trends are observed in the MedBrowseComp-605 results, where we exclusively evaluate models utilizing parametric memories and the RAG framework. The performance of bare models 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.

Table 1: 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.

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 doing RAG, models demonstrate substantial gains on shallow questions (1- and 2-hop) except GPT4.1, Gemini2Flash holds 30% and 4% boost respectively. And 67% and 7.4 % for Gemini2.5Pro. 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.[7] Their 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. Table 1 shows some common error modes in examples.

[7]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, underscoring the urgent need to develop more efficient, unified systems that deliver comparable or superior results with lower resource overhead.

4.2 Computer Use Agent Results

Table 2 presents performance of Anthropic's Computer-Use agent (Sonnet 3.7) on a diagnostic subset of MEDBROWSECOMP. This subset focuses on simpler tasks comparing to the deepresearch partion—each requiring no more than two browser actions—designed to isolate the agent's ability to retrieve and extract specific biomedical fields with minimal navigational complexity. All tasks are grounded in HemOnc.org and span four clinically relevant question types: *Clinical Trial IDs*, *Start Dates*, *Second Authors*, and *PubMed IDs (PMIDs)*.

Extraction Task	**Accuracy (%)**
Clinical Trial IDs	33.88
Start Date	30.58
Second Author	36.36
PMIDs	11.57

Table 2: Anthropic Computer-Use performance on four HemOnc information-extraction tasks in the sandbox environment.

The agent performs best on finding the Second Author of the linked PubMed paper and Trial ID extraction, relatively low-hop tasks that benefit from predictable formatting and shallow navigation. In addition, this information is mostly be found in close proximity on the same page, which likely explains their similar results. Start Date questions require navigating to ClinicalTrials.gov and correctly identifying structured metadata. Errors in these settings often stem from greedy field selection: trajectory analysis shows the agent tends to extract the first date it encounters—even if unrelated to trial start—rather than locating the dedicated "Study Start" field.

As we take one step further beyond Hemonc, asking for the matching PMID - despite appearing structurally simple - results in the weakest performance. Interestingly, this is not due to common visual brittleness or interface failures. Instead, the agent often returns a valid PMID for a plausible but incorrect paper. This suggests semantic confusion rather than surface-level noise. You can see a detailed common error modes sorted with explanations in the following Table 3.

Table 3: Examples of common errors for Sonnet 3.7 on our CUA tasks. We provide three examples, each with attributed common error modes and a detailed explanation screened by the author PM.

Error type & question (paraphrased)	**Explanation**
Type: Greedy attribute extraction **Question:** Give the start date (`YYYY-MM`) for the trial Gemcitabin ± Cisplatin in cholangiocarcinoma.	Agent located the ABC-02 trial but copied "September 2006" from the paper's timeline instead of scrolling down in search of the correct and meaningful start date. The registry lists 2005-05; using a narrative source instead of the structured field produced a one-year error.
Type: Overlooking information **Questions:** Which clinical-trial ID best describes Rituximab monotherapy compared to Ibrutinib monotherapy when used to treat Chronic lymphocytic leukemia.	The model searched and correctly opened the relevant URL (`NCT01234311`). It scrolled through the page but stated it could not find an NCT ID despite it being present, so it continued the search and eventually retrieved `NCT01886872`.
Type: Aimless wondering **Question:** "Give me the second author of the paper about the efficacy of Tranexamic acid monotherapy vs Placebo when used to treat Hereditary hemorrhagic telangiectasia.	The agent correctly opens the relevant NCT trial and attempts to find the associated paper online. However, the first result of the search is not the the actual study. Unable to identify the second author from that source, it explores the linked website for references to any related publications. It then opens the first paper it finds and returns the second author from that unrelated document.

5 Conclusion and Future Work

Our work has certain **limitations**:

System selection: We evaluated only agents with publicly programmable APIs. As a result, promising but closed systems such as OpenAI *Operator* and Google's *Project Mariner* were not evaluated. Other systems like Bytedance's *UITars* were not tested due to time and computational constraints [36]. However, it is not clear that their inclusion would shift the leaderboard substantially given their performance compared to Claude CUA on other benchmarks.

Human supervision: All answers were machine-judged with a lightly-audited LLM rubric. However, a subset of answers were human-verified with good agreement compared to the LLM-as-judge.

Compute and subscription cost: Building the benchmark and running baselines is non-trivial and our experiments cost a total of $\$3,690$: $\$320$ on Perplexity (of which $\$200$ was Deep-Research API calls), $\$450$ on Gemini 2.5 pro, and about $\$2{,}500$ on Claude Sonnet 3.7 CUA, $\$200$ for a 1 month ChatGPT Pro subscription, $\$20$ for advanced reasoning, and $\$200$ for GPT-4.1 mini judging.

Real-world validation: Finally, we did not place answers in front of clinicians , and the questions in MedBrowseComp only cover a small portion of the enormous medical knowledge base used for real-world clinical decision-making. However, expert-curated, verifiable knowledge is not available for the entire medical domain. Our focused benchmark enables deeper study of deep research and computer use capabilities and still reveals important gaps and future directions to inform the field.

5.1 Future Work

Broader task suite: Expand beyond single-field extraction to multi-paragraph justification, guideline concordance, and financial/regulatory trend analysis-all of which require deeper reasoning.

Tool-augmented agents. Test whether lightweight adapters—PDF parsers, table detectors, ClinicalTrials.gov wrappers—allow computer use agents.

Inclusion of closed systems. Collaborate with companies to benchmark Operator, Mariner, and UITars under identical prompts, closing the current comparability gap.

Test system in AI-IDE environment We have not yet placed agents in an AI-IDE sandbox (e.g., Cursor or Windsurf) where they must draft, debug, and reuse their own helper scripts over our horizontal tasks. Such an environment would expose whether a model can sustain state, recover from bugs, and refactor code as requirements evolve.

Study agents with a human-in-the-loop and in real-world settings Future work comparing agentic system vs human performance on MedBrowseComp, and studying top-performing agentic systems in real clinical settings, will be essential to understand the optimal implementation and clinical-translaational role of the benchmark and agentic systems.

5.2 Conclusion

We introduced MEDBROWSECOMP, the largest verifiable benchmark that evaluates deep research and computer use agents in the medical field. The tasks presented in MEDBROWSECOMP are not contrived for difficulty alone; each question mirrors a clinically meaningful information-seeking scenario. Experiments reveal a clear capability gap: retrieval augmented text pipelines answer nearly half as many questions as their deep research counterparts, yet no system, including computer use agents, exceeds 40% accuracy among questions that require more than a single hop.

We find that while GUI-centric agents such as Anthropic's *Computer-Use* demo can control browsers and desktops end-to-end, they tend to under-perform compared to deep researchers driven by APIs on our benchmark. Every GUI action produces roughly two screenshots, inflating the context window to $\sim 200{,}000$ tokens—compared with only $\sim 2{,}000$ tokens for an equivalent text-only workflow. Deepresearch pipelines, by contrast, reach structured endpoints in a single call, eliminating the multi-step visual loop and reducing both latency and cost. MEDBROWSECOMP offers a realistic yet challenging test bed for future deep-research systems and supplies a focused subset that remains demanding—though not prohibitive—for computer-use agents. We hope it will accelerate progress toward efficient, high-accuracy medical question answering across modalities.

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/`.

[36] Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, et al. Ui-tars: Pioneering automated gui interaction with native agents. *arXiv preprint arXiv:2501.12326*, 2025.

Contribution

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 4: 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 accompany with 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.

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 5: 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 tough as you can see from 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. And when you strip away the ability to say "Not applicable" or avoid answering (what we call 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's 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 it struggles. On REAL accuracy for 4-hop questions, it 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 big 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 6: 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 7: 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 Evaluation from Optimal Start Page

To evaluate the impact of structured entrypoints on Computer Use Agent performance, we re-ran the benchmark using the same Claude 3.7 Sonnet system but initialized each task from the HemOnc.org homepage. This strategy avoids ambiguity introduced by external search engines and leverages HemOnc.org's curated structure to simplify task execution.

Table 8: Accuracy (%) of Claude 3.7 Computer-Use Agent on structured extraction tasks, with and without initialization from HemOnc.org homepage. Improvements are shown in bold.

Extraction Task	With HemOnc Start	No Defined Start
PMIDs	**42.98**	11.57
Second Author	**46.28**	36.36
NCT	**39.67**	33.88
Start Date	**31.40**	30.58

Compared to the results reported in Table 4, these scores show notable improvement, particularly on PMID extraction, which increased from 11.57% to 42.98%. By starting from a high-quality structured resource, the agent appears less likely to encounter ambiguous links, irrelevant documents, or noisy intermediate pages.

We hypothesize several contributing factors:

- **Reduced ambiguity at the root**: Starting on HemOnc.org anchors the agent to a semantically dense, domain-verified hub. This prevents early misrouting or unnecessary detours caused by irrelevant or overly generic search results.
- **Fewer hops, higher precision**: Structured navigation from HemOnc often yields answers in one or two clicks. This minimizes the risk of cumulative tool-call errors, selector mismatches, or hallucinated document interpretations.
- **Improved alignment with human workflows**: Clinicians often use HemOnc as a first step for navigating oncology evidence. Our findings suggest that model workflows, like human ones, benefit from strong priors.

These findings suggest that GUI-agent accuracy can benefit significantly from constrained yet clinically meaningful entrypoints like HemOnc.org. Future work should explore formal initialization policies as part of web-agent pipelines.

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.