Demystifying Delays in Reasoning: A Pilot Temporal and Token Analysis of Reasoning Systems

Qi Qi
University of California, San Diego
qiqi@ucsd.edu

Reyna Abhyankar
University of California, San Diego
vabhyank@ucsd.edu

Yiying Zhang
University of California, San Diego
GenseeAI, Inc.
yiying@gensee.ai

Abstract

Despite rapid gains in accuracy, the latency of reasoning and deep-research systems has been largely overlooked. Reasoning models augmented with external tools have demonstrated strong abilities in solving complex tasks. We present the first systematic temporal study of three representative reasoning models and agents, OpenAI o3-deep-research, GPT-5, and the LangChain Deep Research Agent on DeepResearch Bench. By instrumenting each system, we decompose end-to-end request latency and token costs across reasoning, web search, and answer generation. We find that web search often dominates end-to-end request latency and that final answer generation consumes most tokens due to the lengthy retrieved context, implying that tool latency and retrieval design are primary levers for speeding up reasoning end-to-end.

1 Introduction

Reasoning tasks are multi-step problems that require decomposition, intermediate inferences, tool use, and answer writing. They make up numeracy and symbolic math problems [3], open-domain fact finding, software design and debugging [10], and high-stakes decision support in domains like law and medicine. Because of their importance, huge efforts have gone into developing reasoning models [17, 2, 8, 6, 24] and agents [11, 19]. The former are large language models that have built-in reasoning capabilities (*e.g.*, with chain-of-thought [23], self-consistency [22], least-to-most [26], tree-style search [25]) and can optionally trigger external tools [21] such as web search [12], retrieval, calculators, and code runners. The latter are AI agents that package similar capabilities behind orchestration layers for planning, multi-step reasoning, tool usages, and answer synthesis.

Most prior work aims to improve the answer quality/accuracy of reasoning models and agents, with benchmarks like GSM8K [5], MMLU [9], Humanity's Last Exam [18], and DeepResearchBench [7]. Recent discussion has also taken shape on token efficiency and dollar costs of reasoning tasks [20]. However, little attention has been paid on the temporal performance of reasoning tasks, even when many popular AI tasks (*e.g.* computer use) are reported as taking tens of minutes or longer [1].

Going forward, timely reasoning is crucial because many real-world workflows operate under strict latency budgets, where delays degrade outcomes, trust, and user engagement. In on-call incident response and security alert triage, every additional second can widen blast radius and costs. Interactive coding assistants and notebook copilots must respond within a few seconds to preserve the developer or analyst's flow. Customer support and sales assistants in live chats risk abandonment and conversion loss when reasoning lags.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Efficient Reasoning.

In this paper, we conduct (to our knowledge) the first temporal performance study of reasoning models and deep-research agents. We perform detailed evaluation and analysis of three representative reasoning models/agents: OpenAI o3-deep-research [14], OpenAI GPT-5 [16], and the LangChain Deep Research Agent [11], on DeepResearchBench [7], a benchmark that evaluates a language model's ability to generate structured, well-researched, and in-depth analysis on a particular topic. We instrument the pipeline to attribute latency and cost across phases (reasoning, web search/retrieval, final answer generation).

Overall, we found that surprisingly, web search could dominate reasoning task latency, especially for GPT-5 and for long tasks. It accounts for 73% of total wall-clock time on average, overshadowing in-model "thinking." In some cases, web-search can account for up to 91% of end-to-end latency. Moreover, our token and cost analysis shows the final answer generation step consumes the majority of completion tokens across systems. Digging deeper, we found that extensively retrieved context inflates prompts at this stage. These results elevate tool latency and retrieval prompt design, especially web search, from afterthoughts to first-order determinants of user-perceived performance in reasoning workflows. Our instrumentation and analysis framework is open-sourced at https://github.com/WukLab/Deep-Research-Analysis.

2 Study Methodology

DeepResearch Bench [7] is a recent benchmark consisting of 100 PhD-level research tasks crafted by domain experts from 22 fields intended to evaluate deep research agents and reasoning models. A task is typically 1-2 sentences and expected to be answered with a research report. The benchmark evaluates the answer quality based on the comprehensiveness, analysis quality, instruction following, clarity, and citation quality and quantity, using Gemini-2.5-Pro [6] as the judge model. We randomly sampled 10 tasks from the original DeepResearch Bench and grouped them into two sets: five that run longer and five that run shorter.

We choose three representative reasoning models and agents to conduct our pilot study: OpenAI o3-deep-research (o3-DR for short) [14], OpenAI GPT-5 [16], and LangChain Deep Research Agent (LangChain-DR for short) [11]. o3-DR is OpenAI's flagship reasoning model trained specifically for this domain and ranks #2 on the Deep Research Bench Leaderboard [4]. GPT-5 is OpenAI's latest large language model that internally could route to different models and tools depending on task difficulty. LangChain-DR is a popular open-source deep research agent that ranks #6 on the leaderboard (the highest ranking agent). We do not choose the #1 model, Gemini-2.5-Pro, on the leaderboard, because the LLM judge is the same model and can cause biases. Our chosen three systems represent a good variety of top-performing reasoning models and agents; understanding their temporal and token performance can effectively guide future researchers and practitioners in the space.

For each system, we analyze its internal events. For o3-DR and GPT-5, we use OpenAI's response API to capture each internal event, and we categorize them into reasoning, web search, and final answer generation. For LangChain-DR, we instrument the source code to separate each LLM call and tool call into an event but otherwise keep its original source code and default configurations. In our categorization, the "web search" phase includes subsequent summarization or content processing steps performed by the model or agent. For example, in LangChain-DR, this encompasses LLM-based summarization of retrieved pages before reasoning resumes. This ensures consistency across systems, as the OpenAI APIs do not expose finer-grained sub-events within web search.

We use the same prompt template generated by ChatGPT [13] across all systems. For GPT-5, we also analyze the results from its provided low-verbosity setting. Lastly, we use GPT-4.1 [15] as the base model for LangChain-DR as the agentic architecture itself is designed to induce reasoning capabilities. The exact prompt for all models can be found in the Appendix.

3 Study Results

3.1 Cross-System Comparison

A typical task undergoes reasoning intertwined with web search before generating the final output report. We collect the end-to-end latency, the number of tokens for each model stage, and the dollar

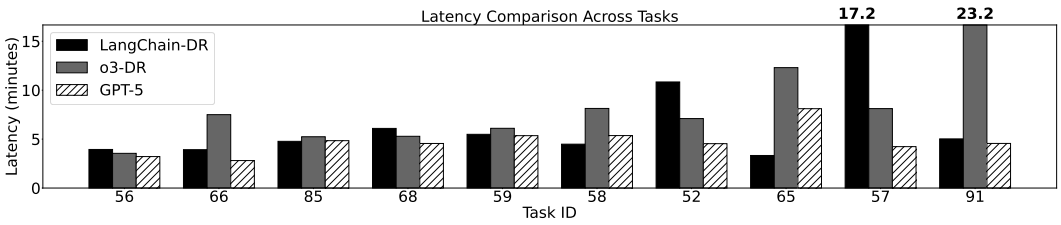

Figure 1: Latency Comparison

Table 1: Latency (min), tokens, cost, and accuracy of different models/agents using long tasks.

		Tokens			
Setting	Latency (min)	Reasoning	Output	Cost ($)	Score
o3-DR	10.52 ± 4.22	4135 ± 1081	15249 ± 5511	1.27 ± 0.26	47.88
GPT-5	5.52 ± 1.37	2241 ± 409	9127 ± 3057	0.28 ± 0.03	47.81
LangChain-DR	18.57 ± 7.72	3527 ± 2692	2147 ± 783	0.57 ± 0.60	40.62

cost, along with the final score in Table 1 and Table 2. For LangChain-DR, we treat all intermediate model steps that do not contribute to the final report generation as “reasoning." Unsurprisingly, o3-DR has the highest reasoning tokens, output tokens, and dollar cost across both long and short tasks. This is due to the API pricing, as well as the specialized nature of the model. Meanwhile, the difference between long and short tasks is most pronounced on LangChain-DR, which is able to achieve 4 points higher accuracy with rougly half the reasoning tokens and cost while operating at one-fifth of the latency. Figure 1 then compares the median latency of 3 trials across different tasks.

Figure 2 illustrates a timeline view of one task in the dataset. The small gaps in the task represent actual gaps in events reported by the OpenAI API. o3-DR and GPT-5 are entirely synchronous, while LangChain-DR has built-in asychronicity (*e.g.*, parallel web search calls followed by calls to a summarizer model). Surprisingly, web search dominates the overall latency for both GPT-5 and LangChain-DR. We explore this phenomenon more thoroughly across all sampled tasks in Figure 4. On both long and short tasks, web-search accounts for a significantly higher share of the overall latency, at 73% on GPT-5 and 50% for LangChain-DR on average.

3.2 System Deep Dive: Verbosity and Web Search

GPT-5 comes with a built-in knob for “verbosity," which allows the user to choose the expressiveness of the model output. We run the sampled workload with the low-verbosity setting and find that the model achieves accuracy on par with the default (medium verbosity) at significantly lower token usage, dollar cost, and latency. The CDFs for these results are displayed in Figure 3. As expected, GPT-5 with lower verbosity is 10% cheaper and faster (due to reduced token usage). However, the corresponding 16.5% drop in accuracy, from 46.9 to 39.2, renders it less efficient on an accuracy-per-dollar basis.

For LangChain-DR, we find the difference in short and long tasks is heavily determined by the number of tokens on the webpage. Figure 5 illustrates that long tasks range from 500K tokens to upwards of 2.5M tokens whereas short tasks range from 100K to 1.25M tokens. This suggests that the tokens produced by web page crawling can impact the end-to-end performance in addition to the

Table 2: Latency (min), tokens, cost, and accuracy of different models/agents using short tasks.

		Tokens			
Setting	Latency (min)	Reasoning	Output	Cost ($)	Score
o3-DR	5.73 ± 1.42	3450 ± 928	6453 ± 2577	0.82 ± 0.26	45.12
GPT-5	3.98 ± 0.82	1818 ± 652	5261 ± 1789	0.19 ± 0.06	46.03
LangChain-DR	4.62 ± 0.83	1966 ± 1061	2327 ± 414	0.26 ± 0.17	44.20

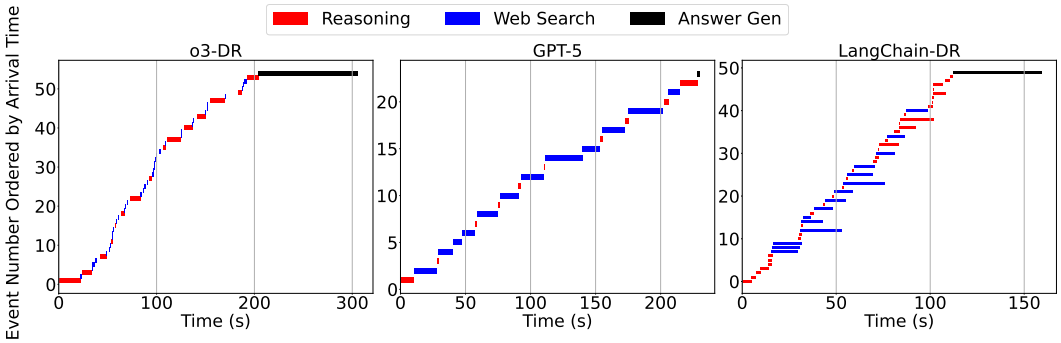

Figure 2: Timeline Comparison

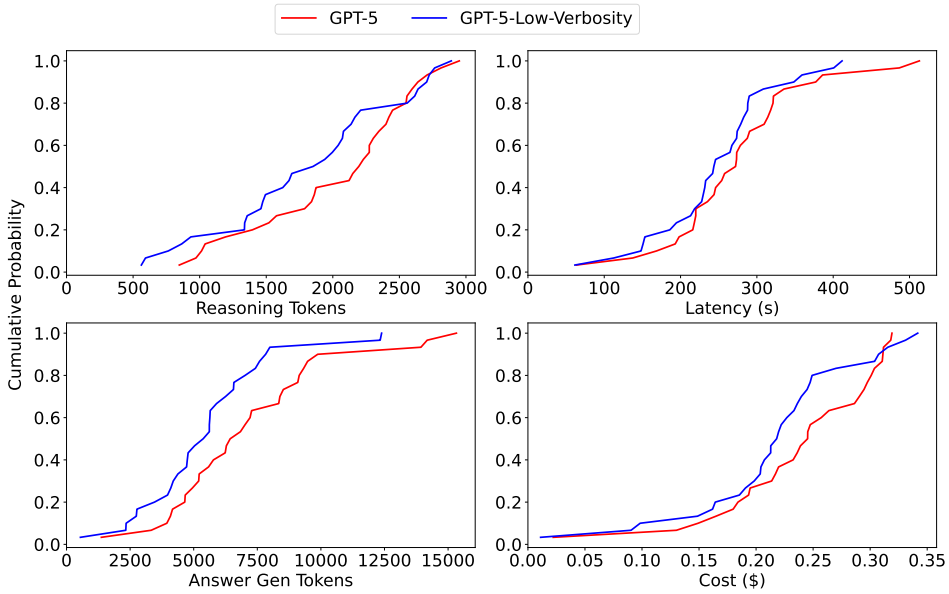

Figure 3: CDFs of GPT-5 and GPT-5-Low-Verbosity

duration of the web search call itself. As seen in Figure 4, LangChain-DR is the only system where shorter tasks spend less time doing web search. Parallelizing web searches or efficient prefetching could potentially reduce latency for long tasks in other systems.

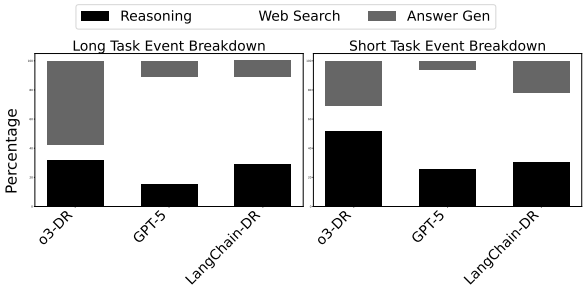

Figure 4: Latency Breakdown by Stage

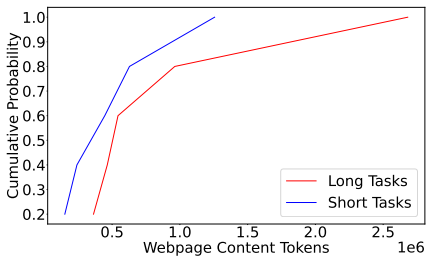

Figure 5: Web Search Tokens CDF by Task Type on LangChain-DR

4 Discussion and Conclusion

Our study highlights that the temporal dynamics of reasoning systems can be shaped heavily by tool latency, particularly web search and have a larger impact than the language models' internal reasoning processes. This holds especially true for generalist reasoning models (*e.g.* GPT-5) and agent architectures that attempt to mimic reasoning. Our pilot study on a subset of Deep Research Bench and systems finds that web search accounts for a significant portion of the end-to-end latency, and this is driven by the tokens outputted by the webpage. These preliminary findings motivate us to conduct a thorough evaluation on the entire benchmark as well as evaluate more reasoning systems. Rethinking tool orchestration can significantly improve model end-to-end latency, which will be critical for real-time workloads that require high levels of reasoning.

Acknowledgment

We would like to thank the anonymous reviewers for their feedback and comments, which have helped improve the content and presentation of this paper. We would also like to thank Zijian He for the valuable contributions and feedback on this paper. This material is based upon work supported by funding from PRISM center (part of SRC's JUMP 2.0), NSF award 2403253, and gifts from AWS, Google, and Meta. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of these institutions.

References

[1] Reyna Abhyankar, Qi Qi, and Yiying Zhang. Osworld-human: Benchmarking the efficiency of computer-use agents. *arXiv preprint arXiv:2506.16042*, 2025.

[2] Anthropic. Claude opus 4.1. `https://www.anthropic.com/claude/opus`, 2025. Accessed: 2025-09-08.

[3] Art of Problem Solving Wiki. AIME Problems and Solutions, 2025. Accessed 2025-09-08.

[4] Mingxuan Du (Ayanami0730). Deepresearch bench leaderboard, 2025. Accessed: 2025-09-08.

[5] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Łukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. Introduces GSM8K benchmark.

[6] Google DeepMind. Gemini 2.5: Our most intelligent ai model. `https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/`, 2025. Accessed: 2025-09-08.

[7] Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. Deepresearch bench: A comprehensive benchmark for deep research agents, 2025.

[8] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[9] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[10] Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. SWE-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations*, 2024.

[11] LangChain-AI. open_deep_research. `https://github.com/langchain-ai/open_deep_research`, 2025. Accessed: 2025-09-08.

[12] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

[13] OpenAI. Introducing chatgpt, 2025. Accessed: 2025-09-08.

[14] OpenAI. Introducing deep research, 2025. Accessed: 2025-09-08.

[15] OpenAI. Introducing gpt-4.1 in the api, 2025. Accessed: 2025-09-08.

[16] OpenAI. Introducing gpt-5, 2025. Accessed: 2025-09-08.

[17] OpenAI. Introducing openai o3 and o4-mini. https://openai.com/index/introducing-o3-and-o4-mini/, 2025. Accessed: 2025-09-08.

[18] Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.

[19] Pranav Putta, Edmund Mills, Naman Garg, Sumeet Motwani, Chelsea Finn, Divyansh Garg, and Rafael Rafailov. Agent q: Advanced reasoning and learning for autonomous ai agents. *arXiv preprint arXiv:2408.07199*, 2024.

[20] Nous Research. Measuring thinking efficiency in reasoning models: The missing benchmark. https://nousresearch.com/measuring-thinking-efficiency-in-reasoning-models-the-missing-benchmark/, 2025. Accessed: 2025-09-08.

[21] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*, 2023.

[22] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

[23] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

[24] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[25] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023.

[26] Denny Zhou, Nathanael Schärli, Hou Le, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*, 2022.

A Appendix

A.1 Prompt Template

The following is the prompt template used as the system prompt for o3-deep-research and LangChain-DR and the developer prompt for GPT-5:

> Deep Research on user's query:
>
> Objectives: - Provide a rigorous, data-backed analysis. - Include concrete facts, recent trends, and measurable impacts. - Prioritize authoritative, up-to-date sources. - Use inline citations (e.g., [Author Year] or [Org Year]) and return full source metadata (title, org, URL, date). - Be explicit about assumptions, uncertainty, and data limitations.
>
> Method: 1) Use web search to find and cross-verify >= 6 sources (>= 3 primary). 2) Include sensitivity notes or policy.
>
> Deliverables: - Executive summary (bulleted, 5-8 key findings). - Analysis. - Sensitivity notes and uncertainties. - Source list with full metadata and citation keys used in text.
>
> Constraints: - Analytical tone; avoid generalities. - No hallucinated data; every quantitative claim must have a citation.