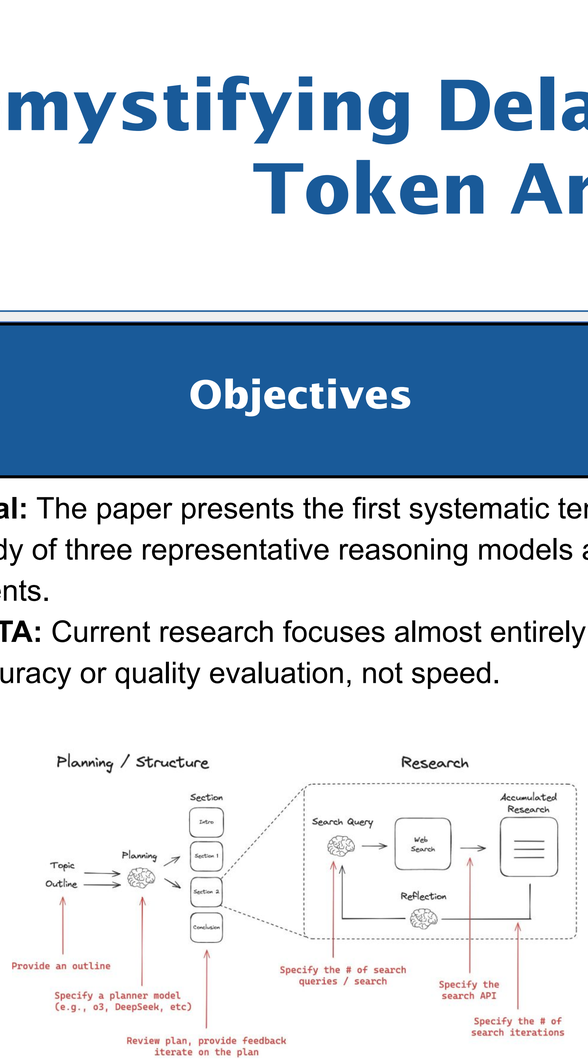
Theme 1 - 3135.003

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

Q.Qi (MS), R. Abhyankar (PhD); Y. Zhang (PI@UCSD)

Objectives

- **Goal:** The paper presents the first systematic temporal study of three representative reasoning models and agents.
- **SOTA:** Current research focuses almost entirely on the accuracy or quality evaluation, not speed.



- **Challenges:** Existing top-tier deep research frameworks conduct tasks by running complex workflows with tool-callings, making it difficult to find the exact bottlenecks with a unified standard.
- **Objective**:
  - To highlight that the temporal dynamics of reasoning systems are heavily shaped by tool latency, particularly web search, often more than the language models' internal reasoning processes.
  - To motivate rethinking tool orchestration to improve end-to-end latency, which will be critical for real-time workloads requiring high levels of reasoning.

Technical Approach

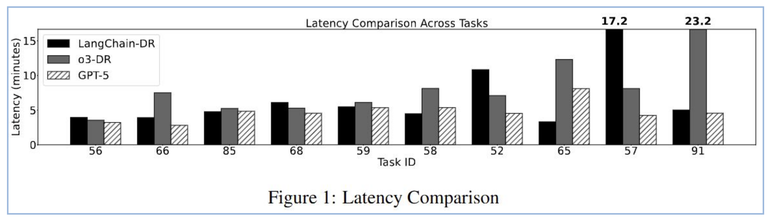


Figure 1: Latency Comparison

- For O3-DR and GPT-5, the OpenAI response API was used to capture and categorize internal events into reasoning, web search, and final answer generation. For LangChain-DR, the source code was instrumented to separate each LLM call and tool call into an event.
- A typical task involves reasoning intertwined with web search before generating the final output report.
- The "web search" phase was defined to include subsequent summarization or content processing steps to ensure consistency across systems.
- O3-DR and GPT-5 are entirely synchronous, while LangChain-DR has built-in asynchronicity, such as parallel web search calls.

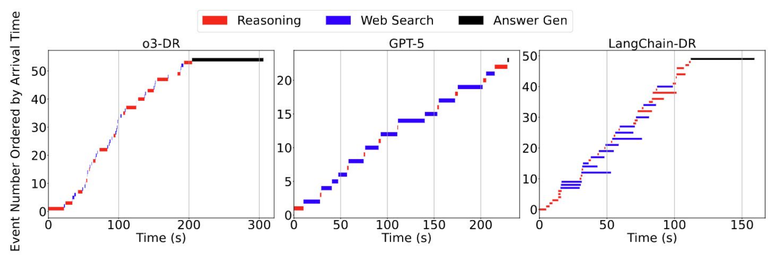


Figure 2: Timeline Comparison

Technical Approach

- GPT-5 Verbosity: low-verbosity setting was tested and found to be 10% cheaper and faster with a 16.5% drop in accuracy and less efficient on an accuracy-per-dollar basis.

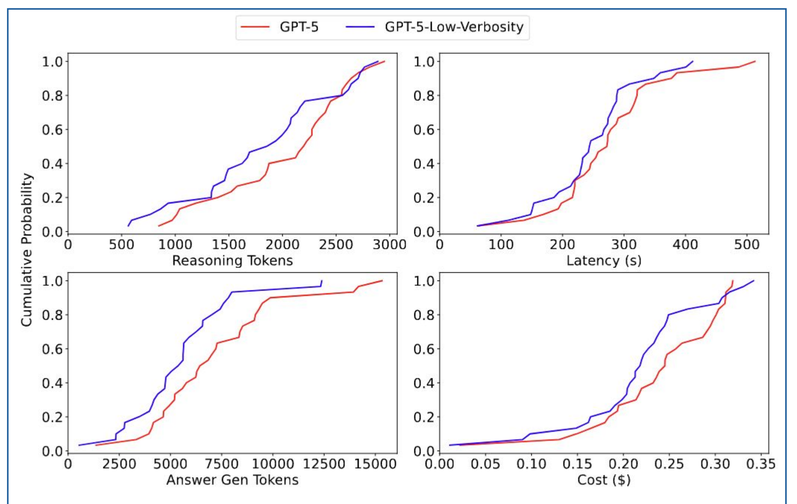


Figure 3: CDFs of GPT-5 and GPT-5-Low-Verbosity

- LangChain-DR Web Search:
  - The difference between short and long tasks was found to be heavily determined by the number of tokens on the webpage.
  - This suggests that the tokens produced by web page crawling can impact end-to-end performance.

Experimental Setup

- **Systems**: OpenAI O3-deep-research (O3-DR), OpenAI GPT-5, and LangChain Deep Research Agent (LangChain-DR).
- **Benchmark**: 10 tasks (5 long, 5 short) randomly sampled from DeepResearch Bench.
- **Evaluation**: The benchmark uses Gemini-2.5-Pro as the judge model to evaluate answer quality based on comprehensiveness, analysis quality, and other factors.
- **Metrics**: End-to-end latency, tokens for each stage (reasoning, output), dollar cost, and final accuracy score.

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

Setting	Latency (min)	Tokens		Cost ($)	Score
		Reasoning	Output		
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

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

Setting	Latency (min)	Tokens		Cost ($)	Score
		Reasoning	Output		
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

Results and Comparison to the State-of-the-Art

- Key Finding: Web search, not in-model "thinking," dominates the overall latency for both GPT-5 and LangChain-DR.
- On average, web search accounts for 73% of total wall-clock time on GPT-5 and 50% for LangChain-DR. In some cases, web search can account for up to 91% of end-to-end latency.
- The final answer generation step consumes the majority of completion tokens across all systems. This is due to extensively retrieved context inflating the prompts at this stage.

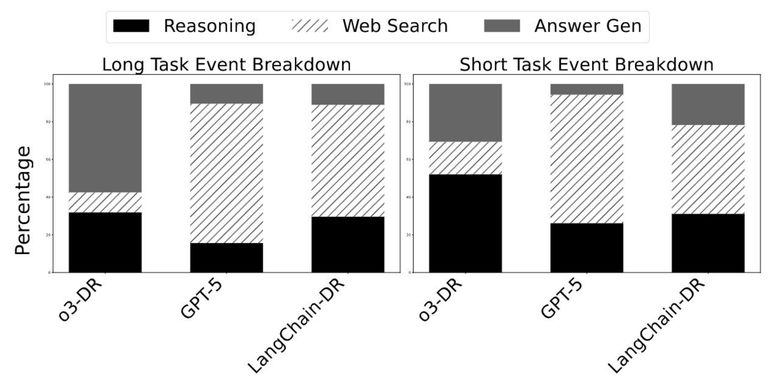


Figure 4: Latency Breakdown by Stage

Key Accomplishments, Showstoppers & Next Steps

- **Key accomplishments:**
  - Identified that tool latency (specifically web search) and retrieval design are primary levers for speeding up reasoning end-to-end.
  - Found that web search can dominate end-to-end request latency , accounting for 73% of total time on average for GPT-5.
  - Accepted to the Workshop on Efficient Reasoning at NeurIPS 2025.
- **Grand Challenge Applications**
  - Efficient reasoning can potentially be leveraged during both deep insight and drug discovery.
- **Demo**
  - The instrumentation and analysis framework used in this study is open-sourced at: https://github.com/WukLab/Deep-Research-Analysis .
- **Lessons learned:**
  - We learned the primary bottleneck is not the model's reasoning time, but the surprisingly high latency of its tools, with web search dominating up to 91% of the total request time.
- **Showstoppers:**
  - None.
- **Next steps:**
  - Explore mechanisms to improve the temporal efficiency of reasoning models and deep research agents.

This work was supported in part by the Semiconductor Research Corporation (SRC) and DARPA.

Publications:

Publication: **Qi Qi**, Reyna Abhyankar, Yiying Zhang, “Demystifying Delays in Reasoning: A Pilot Temporal and Token Analysis of Reasoning Systems,” the 1st Workshop on Efficient Reasoning Co-Located with NeurIPS 2025 (ER '25);
Publication: Reyna Abhyankar, **Qi Qi**, Yiying Zhang, “OSWorld-Human: Benchmarking the Efficiency of Computer-Use Agents,” the 1st Workshop on Computer-Use Agents Co-Located with ICML 2025 (WUCA '25);

AVAILABLE FOR HIRE

Semiconductor Research Corporation

PRISM Annual Review | November 12-13, 2025

UC San Diego · Georgia Tech · The University of Texas at Austin · University of Illinois Urbana-Champaign · UCLA · Penn State · Stanford University · University of Virginia · UCI · University of Wisconsin-Madison · University of Washington · Berkeley University of California

PRISM