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

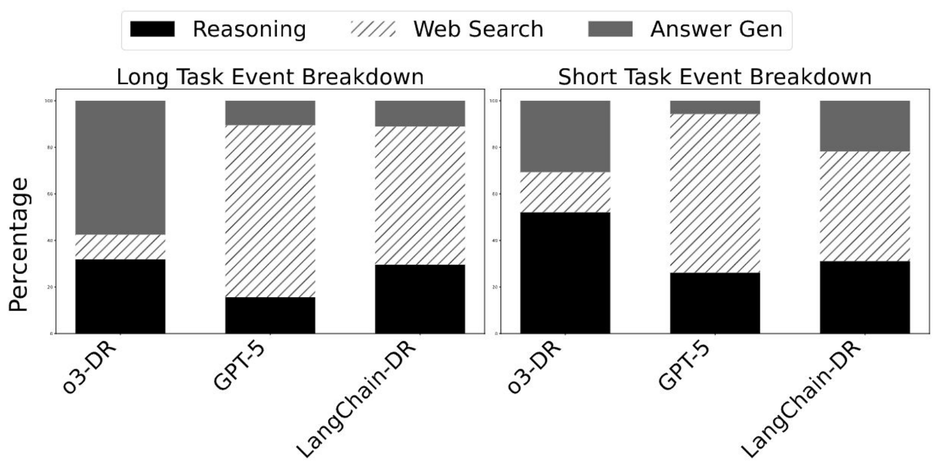
Q.Qi (MS),
R. Abhyankar (PhD);
Y. Zhang (PI@UCSD)

Paper

Blog

Key Findings

- **Web search** often dominates **e2e** latency.
- **Parallelism** and **asynchronous** tool execution helps.



Latency Breakdown by Stage

Methodology

Systems:

OpenAI O3-deep-research, OpenAI GPT-5, and LangChain Deep Research Agent.

Workloads:

10 tasks (5 long, 5 short) sampled from the **DeepResearch Bench**.

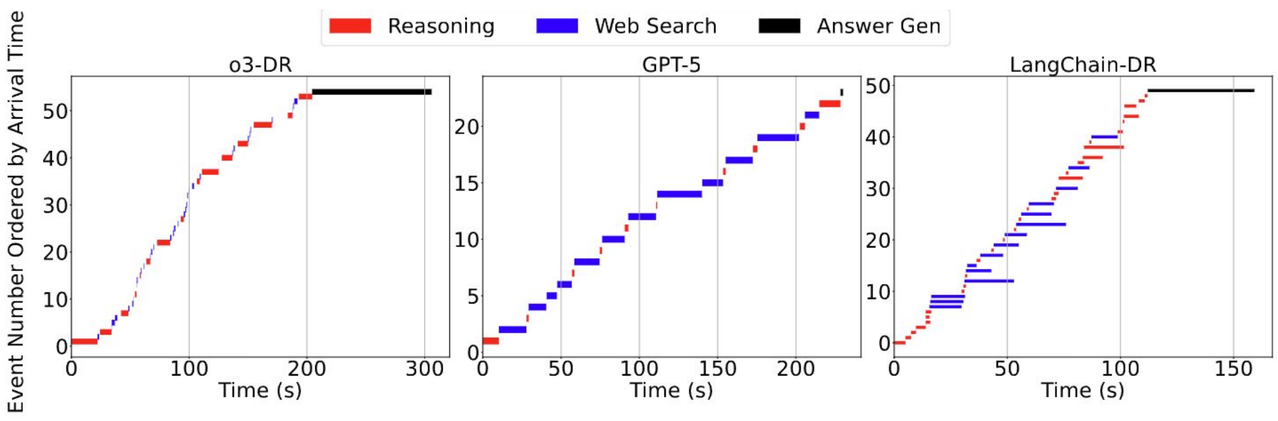
Profile Metrics:

End-to-end latency, tokens for each stage (reasoning, output), dollar cost, and final accuracy score.

Future Work

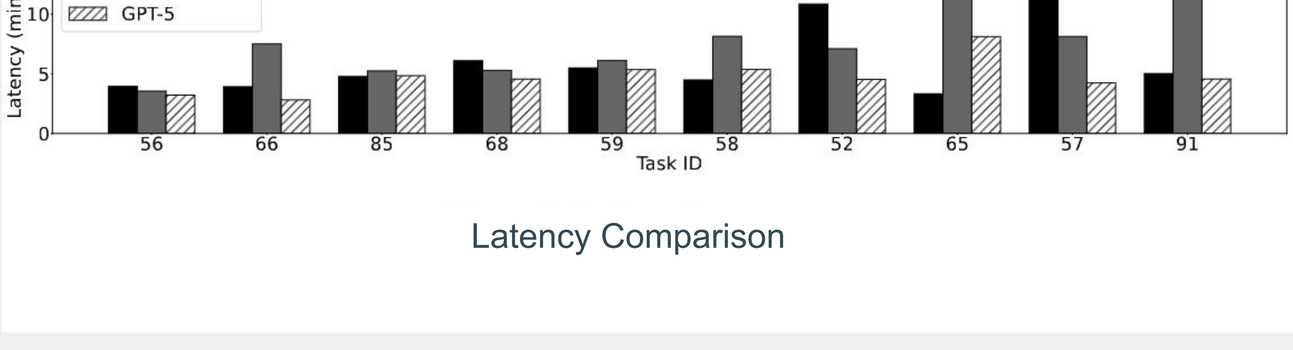
- Study with full DR and other benchmarks.
- Tool acceleration.
- Parallel and asynchronous tool execution.
- Tool speculation.
- More concise and effective retrieval.
- Better agent-tool interaction.

Study Results - Latency Breakdown



Latency Breakdown by Stage

- 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 **answer generation** step consumes the **majority** of **completion tokens** across all systems, due to extensively **retrieved context** inflating.



Latency Comparison

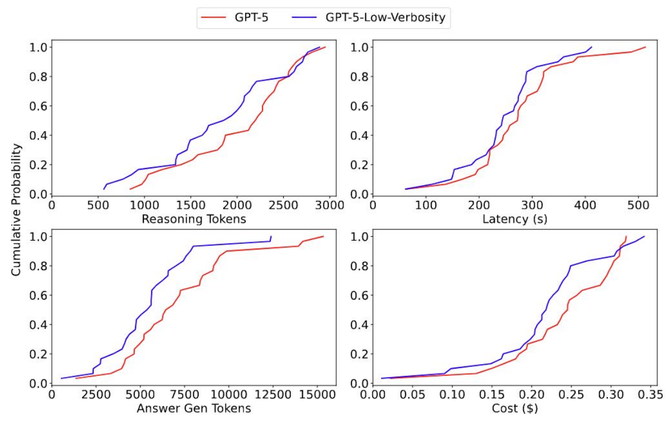
Study Results - Token Analysis

Latency (min), tokens, cost, and accuracy of different models/agents using **long tasks**

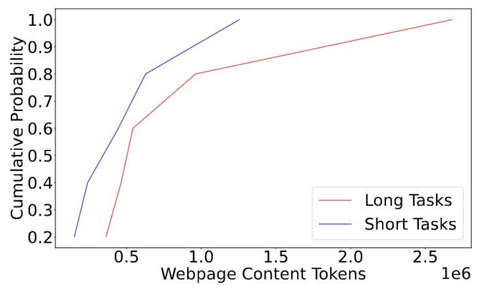
Setting	Latency (min)	Tokens		Cost ($)	Score
		Reasoning	Output		
o3-DR	10.52 ± 4.22	4135 ± 1081	15249 ± 5511	1.27 ± 0.26	47.88
GPT-5	5.52 ± 1.37	2241 ± 409	9127 ± 3057	0.28 ± 0.03	47.81
LangChain-DR	18.57 ± 7.72	3527 ± 2692	2147 ± 783	0.57 ± 0.60	40.62

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



CDFs of GPT-5 and GPT-5-Low-Verbosity



Web Search Tokens CDF by Task Type on LangChain-DR

GPT-5 Verbosity:

low-verbosity found to be **10% cheaper** & **faster** with a **16.5% drop in accuracy**.

LangChain-DR Web Search:

The **short/long**-task-difference is determined by the **# tokens** of **webpages**.