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

Presenter: Qi Qi (UCSD), Reyna Abhyankar (UCSD)
Yiying Zhang (UCSD)
Project #: T1 - 3135.003

- **Problem & State of the Art**
  - While AI reasoning systems are getting more accurate, their latency has been "largely overlooked". Reasoning tasks such as deep research are complex.
- **Technical Approach**:
  - 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.
  - By identifying that tool latency (not model thinking) is the main bottleneck, we provide a clear direction for optimizing the end-to-end performance and efficiency of these critical systems.
- **Results & metrics**:
  - Key findings: Web search dominates latency, answer generation dominates token costs.
  - Metrics: End-to-end latency, tokens per stage, dollar cost, and final accuracy score.
- **Showstopper**
  - None.
- **Grand challenge application & demo**:
  - Efficient reasoning can potentially be leveraged during both deep insight and drug discovery.

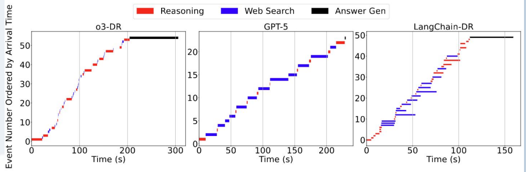

Figure 2: Timeline Comparison

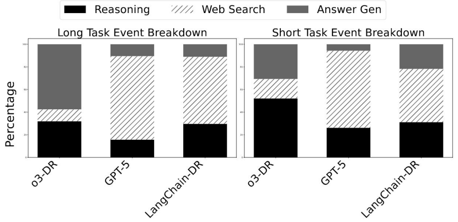

Figure 4: Latency Breakdown by Stage

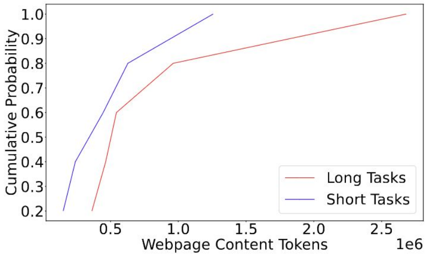

Figure 5: Web Search Tokens CDF by Task Type on LangChain-DR

Github: https://github.com/WukLab/Deep-Research-Analysis.

Publication: 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).