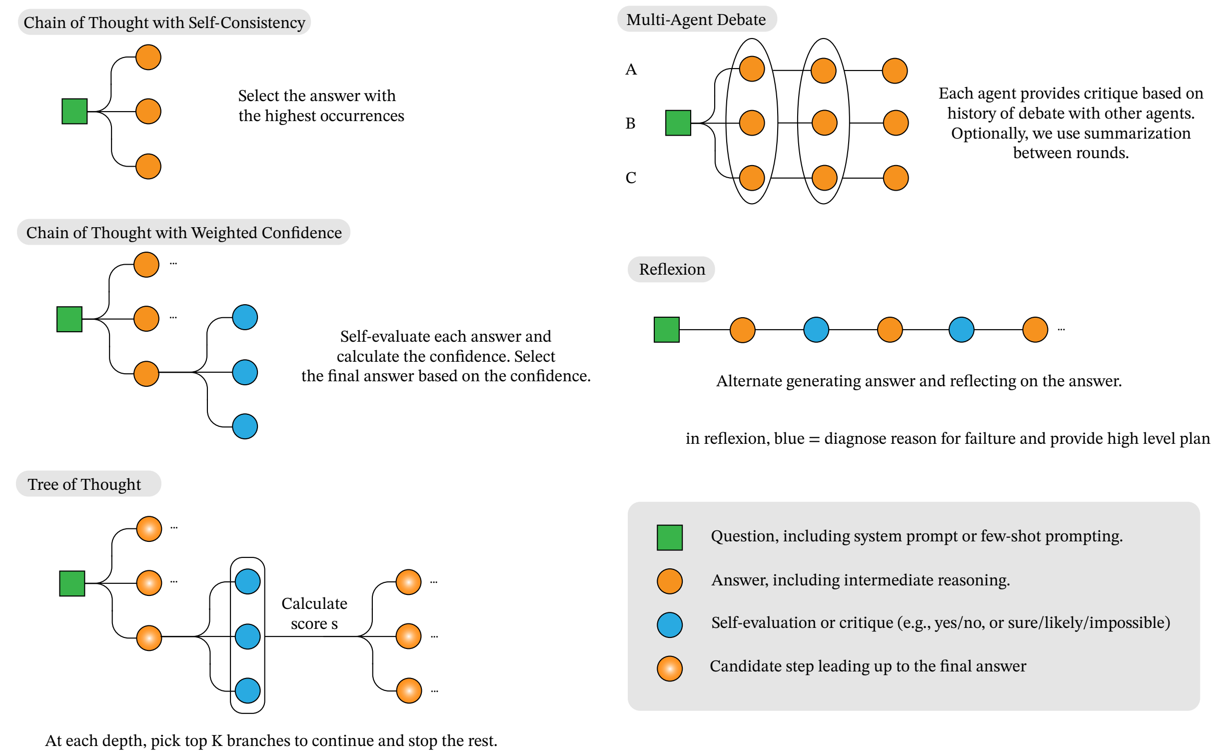
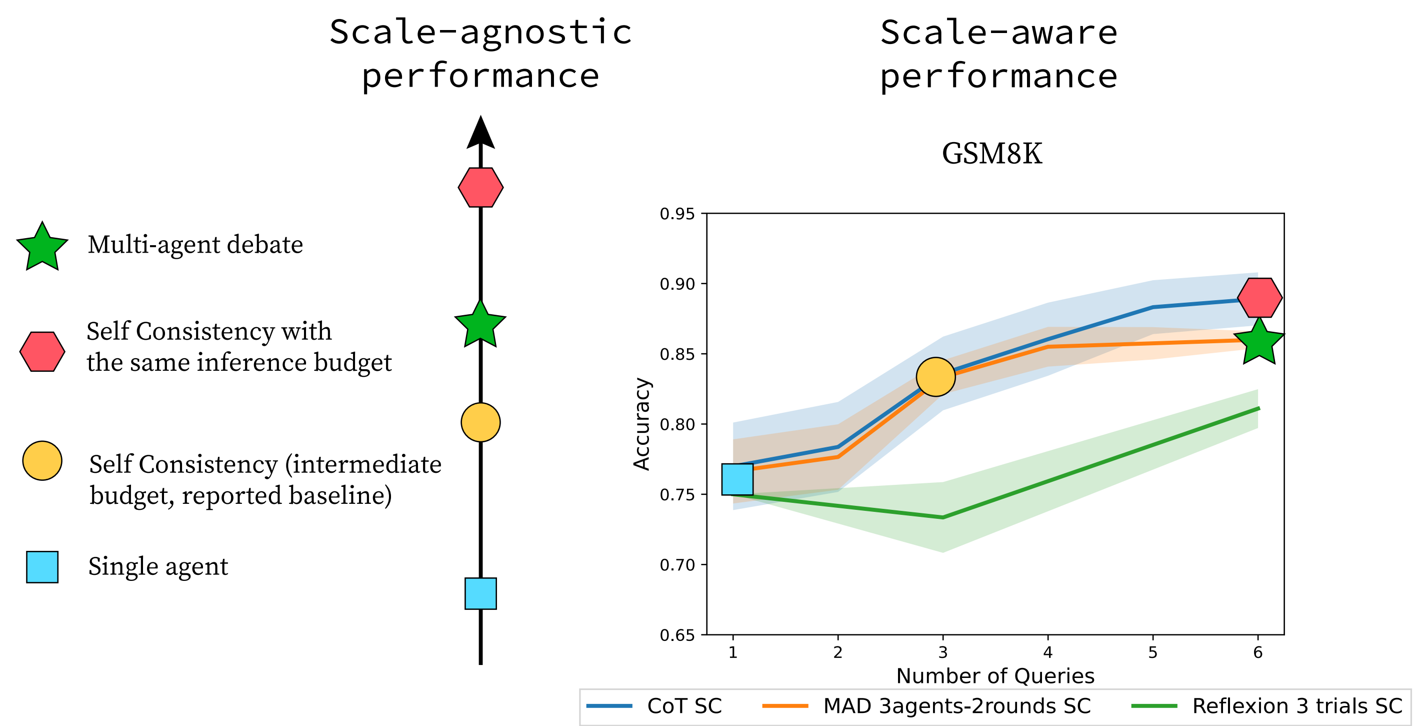
Reasoning in Token Economies: Budget-Aware Evaluation of LLM Reasoning Strategies

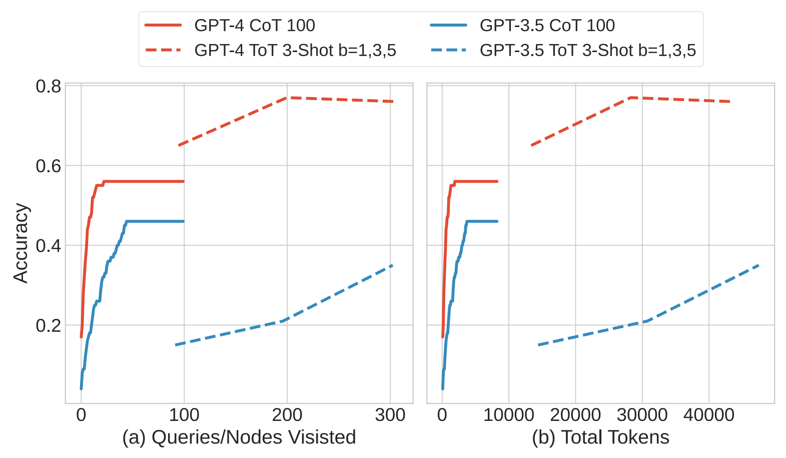
Junlin Wang, Siddhartha Jain, Dejiao Zhang, Baishakhi Ray, Varun Kumar, Ben Athiwaratkun
1) Duke University 2) AWS AI

Overview





Evaluation on Tree of Thoughts

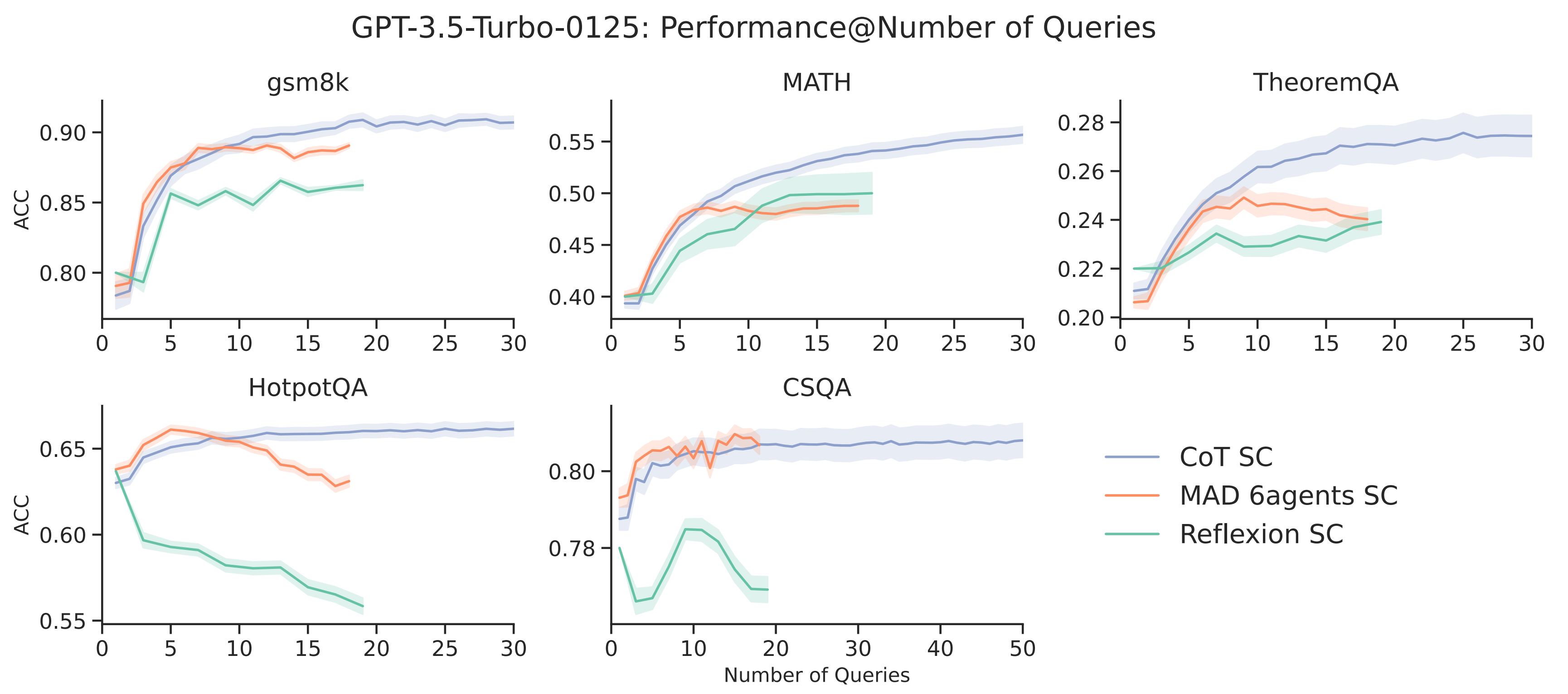


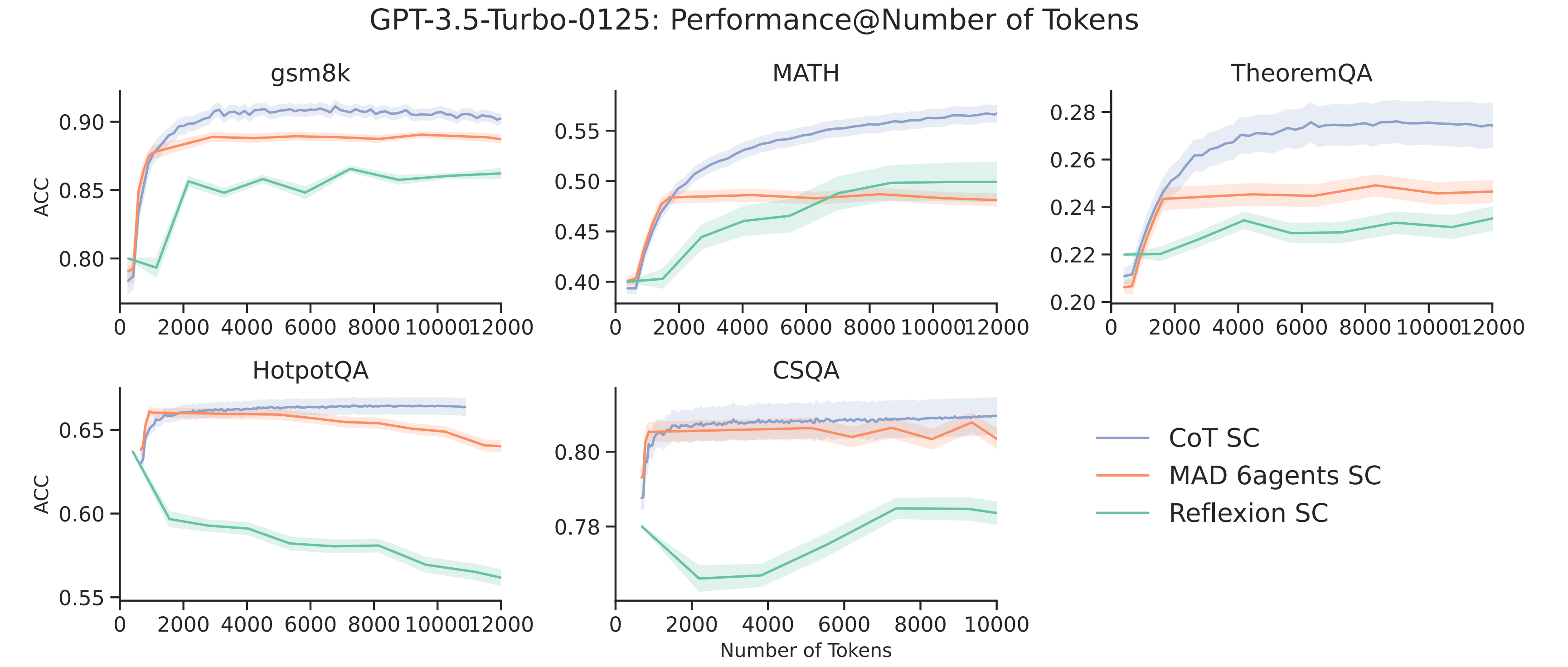
On GPT-4 Tree-of-thoughts beats CoT SC by a big margin but requires way more tokens. On weaker model like GPT-3.5, simpler strategy like CoT beats Tree-of-thoughts by a considerable margin.

Contributions

- Our analysis reveals that traditional evaluation metrics often overlook a critical aspect: the performance gains achievable through additional computational resources. This observation is strongly supported by our comprehensive comparison of Chain-of-Thought Self-Consistency (CoT SC), where we demonstrated that CoT SC not only competes but often surpasses more complex reasoning strategies in effectiveness.
- We introduced a budget-conscious evaluation framework spanning three dimensions: queries, tokens, and monetary cost.
- Furthermore, we investigated the influence of two budget types—generation and evaluation—on the Tree of Thought (ToT) methodology. Our findings highlight that its advantages become more significant with advanced models like GPT-4, partly because of GPT-4's superior evaluation performance.

Evaluating Reasoning Strategies with Budget-Aware Evaluations

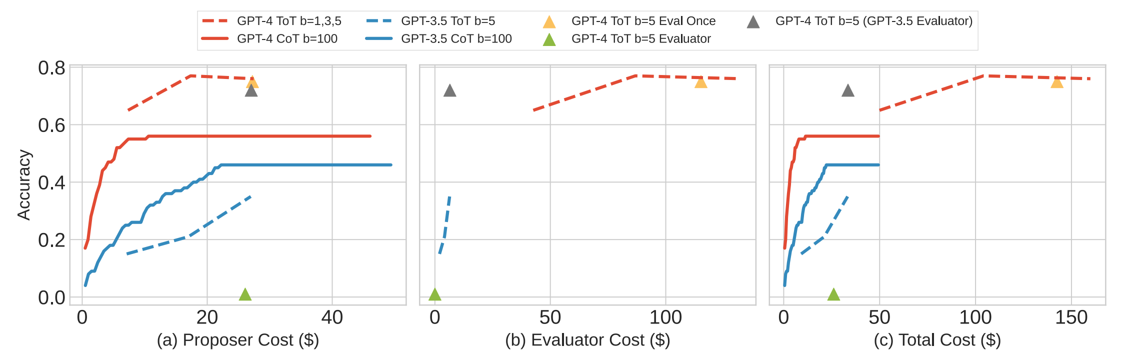




Budget Definitions

1. **API Monetary cost** is generally represented as $c = \alpha_1 * n_I + \alpha_2 * n_O$. Here, n corresponds to the number of input and output tokens. The coefficients are specific to the LLM API in use.
2. **Total number of tokens**, a straightforward metric, is described by $t = n_I + n_O$.
3. **Number of queries** of planned API calls can be a rough proxy for the budget.

Separating Generation/Evaluation Budget



Insights from our method:

- if we use a weaker evaluator like GPT-3.5, we can maintain most of the performance while being very cost-efficient.
- Quantify how much impact each component has: answer generation vs. self evaluation