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

Junlin Wang[1*], **Siddhartha Jain** [2], **Ben Athiwaratkun** [2], **Dejiao Zhang** [2], **Baishakhi Ray** [2], **Varun Kumar** [2]

[1]Duke University [2]AWS AI Labs

Abstract

A diverse array of reasoning strategies has been proposed to leverage the capabilities of Large Language Models (LLMs). In this paper, we point out that traditional evaluations that focus solely on performance metrics miss a key factor: the increased effectiveness due to additional compute. By overlooking this aspect, a skewed view of strategy efficiency is often presented. This paper introduces a framework that incorporates the compute budget into the evaluation, providing a more informative comparison that takes into account both performance metrics and computational cost. We find that a complex reasoning strategy does not always benefit unconditionally from scale, but the performance can plateau quickly. A baseline approach such as Self-Consistency (SC) or majority vote, although algorithmically very simple, can often scale with respect to compute better than complicated strategies and hence perform better overall. These findings shed light on how the budget-aware framework should be used to compare various reasoning strategies, which may spur the development of more robust and cost-effective reasoning strategies and LLM applications.

1 Introduction

The arena of large language models (LLMs) such as GPT-4 (OpenAI, 2023) has seen a proliferation of diverse reasoning strategies. However, comparing these strategies fairly and comprehensively has proven to be a challenging task due to their varied computational requirements. For instance, strategies like the Tree of Thoughts (ToT) necessitate branching out into multiple sequences and incorporating self-evaluation, making them more compute-intensive than others. Therefore, an evaluation framework that only accounts for performance metrics may miss crucial practical factors such as computational cost.

In this paper, we propose the inclusion of the compute budget into the performance measurement of different reasoning strategies. This budget-aware comparison yields a more balanced perspective on the effectiveness of reasoning strategies, accounting for both the quality of the output and the computational resources expended.

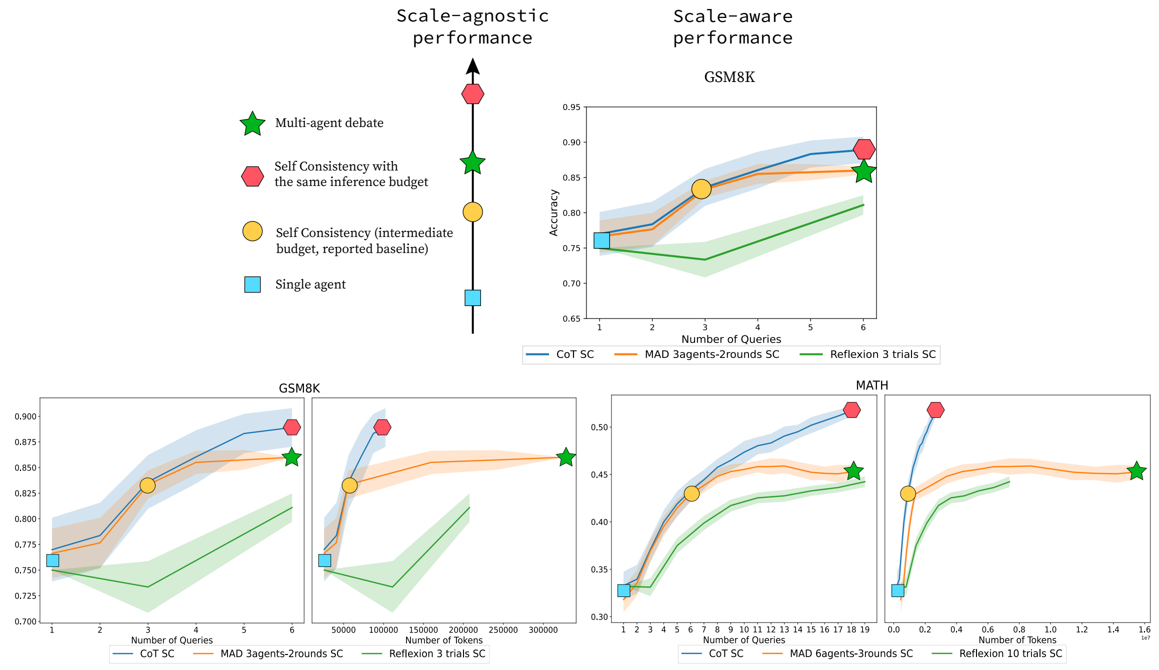
Concretely, our contributions are

- We present a comprehensive head-to-head evaluation of multiple LLM reasoning strategies on multiple types of datasets using GPT-3.5, and GPT-4.
- We evaluate the performance of the strategies on a novel dimension – performance w.r.t. budget. Specifically, we propose 3 types of metrics: performance@number of queries, performance@number of tokens, and perfomance@monetary cost and find that SC is the strongest compared to all other strategies for most models and datasets except for ToT with GPT-4.
- We show that reasoning strategies can benefit from inference scales differently. For instance, some complex strategies such as multi-agent debate encounter performance plateau after the second round of debate.

2 Inference Budget of Reasoning Strategies

While the raw performance of different prompting or reasoning strategies for LLMs is a common topic, how different strategies perform when *budget-constrained* is less well-studied (with the notable exception of Olausson et al. (2023)). However taking budget into account can be critical when using LLMs. In this section we describe different usage scenarios that a user could be interested in and what budgetary metrics would be relevant to those scenarios.

[*]Work conducted during an internship at Amazon



(a) GSM8K

(b) MATH

Figure 1: (1) Comparison of reasoning approaches multi-agent debate (MAD) against the SC baseline, considering both scale-agnostic and scale-aware evaluation, with published scores and our reproductions on the GSM8K dataset. The scale-aware evaluation furnishes more comprehensive insights into the influence of scale on reasoning strategies and offers a fairer method of comparison. (2) The scale-aware comparison between Reflexion and SC on HotpotQA also illustrates the artifact of scale on performance. In both (a) and (b), we show both budgets, the number of total tokens, and the number of queries. All results were obtained from GPT-3.5.

2.1 Budget

We examine various budgetary metrics for LLMs. Given that the number of input and output tokens often feature prominently across these metrics, we designate them as n_I and n_O respectively.

i) **API Monetary cost** is generally represented as $c = \alpha_1 \cdot n_I + \alpha_2 \cdot n_O$. Here, n_I and n_O correspond to the number of input and output tokens. The coefficients α_1 and α_2 are specific to the LLM API in use.

ii) **Total number of tokens**, a straightforward metric, is described by $t = n_I + n_O$. This becomes pertinent when $\alpha_1 = \alpha_2$, which is true for many LLM APIs and is also reflective of the compute cost. Its simplicity ensures it doesn't inherently favor any specific model or API provider.

iii) **Number of queries** of planned API calls can a rough proxy for budget. Such number can be determined before inference, which can give us a rough guidance before actually performing each reasoning strategies. Note that in case we want to sample multiple outputs from the LLM, we count those as *separate* queries

3 An Evaluation in Budget-Aware Metrics

3.1 Importance of Inference Scale

Results in Figure 1 elucidate the efficacy of reasoning techniques, including multi-agent debate and Reflexion, in contrast with the SC baseline. The SC baseline regularly outperforms more complex strategies when given equivalent budgets. Relying solely on scale-independent assessments, as is sometimes done in prior works, might lead to incomplete or potentially misleading interpretations.

3.2 Does a higher inference budget always lead to better reasoning?

As seen in Figure 1b, we find that the SC baseline exhibits a smooth increase in scores with respect to scale. However, such a trend does not always hold for other reasoning strategies. For instance, in multi-agent debate (MAD), an augmented inference budget eventually experiences a performance plateau.

References

Theo X Olausson, Jeevana Priya Inala, Chenglong Wang, Jianfeng Gao, and Armando Solar-Lezama. 2023. Demystifying gpt self-repair for code generation. *arXiv preprint arXiv:2306.09896*.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.