Large Language Model Cascades with Mixture of Thought Representations for Cost-Efficient Reasoning

Murong Yue[1], **Jie Zhao**[2], **Min Zhang**[3], **Liang Du**[2], **Ziyu Yao**[1]
[1]George Mason University [2]Microsoft [3]Virginia Tech
[1]{myue, ziyuyao}@gmu.edu [2]{zhaojie, liang.du}@microsoft.com
[3]{minzhang23}@vt.edu

1 Introduction

Large language models (LLMs) such as GPT-4 have exhibited remarkable performance in reasoning tasks (Rae et al., 2021; Lewkowycz et al., 2022; Zhong et al., 2023). Because of the intensive computing resources required for LLM, we are motivated to study strategies for reducing the costs of using LLMs while not sacrificing task performance, particularly for LLMs' applications to reasoning tasks. Our intuition is that simple questions could be answered by the weaker but more affordable LLM, whereas only the difficult questions need to be tackled by the more expensive, stronger LLM.

Chen et al. (2023) explored the idea of "LLM cascades", where a question is always first answered by a weaker LLM, and then optionally routed to a stronger LLM when the the weaker LLM's answer is not accepted. To decide this routing, this work suggested fine-tuning a smaller LLM to score each question along with its answer produced by the weaker LLM. While this approach could work for some tasks, in practice, we observed that it did not yield satisfying performance for intricate reasoning tasks. Intuitively, it is very challenging to evaluate the difficulty and the answer correctness of a reasoning question solely based on its literal expression, even with a large enough LLM, since the errors could be nuanced despite the reasoning paths appearing promising (Madaan et al., 2023).

In this work, we proposed to devise this routing decision-maker from a different angle, i.e., the "answer consistency" of the weaker LLM (Wang et al., 2023). In particular, we proposed to leverage a "mixture of thought (MoT) representations", which samples answers from both Chain-of-Thought (Wei et al., 2022, CoT) and Program-of-Thought (Chen et al., 2022; Gao et al., 2023, PoT) prompts, emulating how experts can provide diverse perspectives to the same question. Our approaches based on a mixture of thought representations achieved comparable task performance with only 40% of the cost of GPT-4.

2 LLM Cascades for Cost-Efficient Reasoning

The core of our LLM cascade is the decision maker, which takes in the output from the weaker LLM, and then decides whether to route to the stronger LLM or not. We propose two methodologies based on the "answer consistency" of the weaker LLM.

Answer Consistency and Sources of Sampling Answer consistency has been found helpful for improving the LLM performance in reasoning tasks (Wang et al., 2023). Drawing inspiration from prior works, we make the following hypothesis: When the weaker LLM samples highly consistent answers for a given question, it reveals a high "confidence" in solving this question. In this case, there is thus no need to invoke the stronger LLM. While existing literature typically investigated either CoT or PoT independently, in this work, we propose to leverage the synergy of both thought representations in a single task. We hypothesize that an LLM obtains truly high confidence in its problem-solving, only when it is able to produce a consistent answer agnostic to how the intermediate steps are represented. Therefore, we propose to sample the weaker LLM answers from a "mixture of thought (MoT) representations", which includes both CoT and PoT prompts.

Method 1: Vote-based decision-making The first method calculates the consistency of the weaker LLM's answer samples by voting. The most consistent answer can be selected as the one that most samples agree with, and this answer will also be regarded as the final answer A^w by the weaker LLM. The decision maker measures the weaker LLM's consistency via the agreement score. The larger the score is, the more consistent the

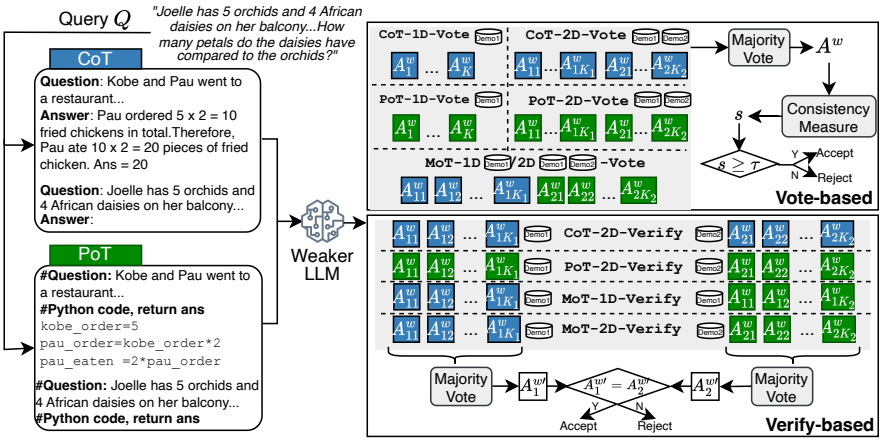

Figure 1: An overview of our approaches (6 vote-based and 4 verification-based). We use ■ to represent the answers from PoT and ■ to represent the answers from CoT. $Demo_i$ is the i-th set of demonstrations.

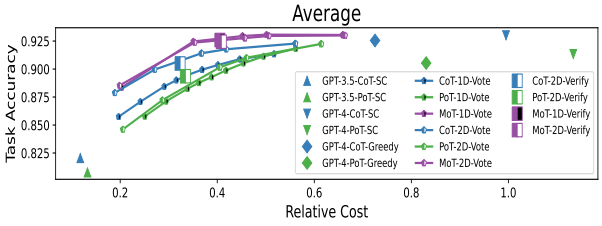

Figure 2: The average performance over 6 reasoning datasets.

weaker LLM's answer samples. In conjunction with a pre-defined threshold value, the decision maker accepts the weaker LLM's most consistent answer A^w when the agreement score is higher than the threshold and rejects it otherwise. As a result, the total cost of answering a question can vary depending on the threshold.

Method 2: Verification-based decision-making In the case of producing samples from two different prompt settings (i.e., different demonstrations or thought representations), we propose the second method, which compares the most consistent answers produced by each prompt. Our method verifies the most consistent answers within each prompt. Only when the two answers are the same, the weaker LLM's answer will be accepted by the decision maker. In this case, the final answer of the weaker LLM will be the same as the two most consistent answers.

3 Experiment

The details of our experiment are shown in Appendix A. The conclusions are:

- Our pipeline achieves comparable task performance with significantly reduced costs.On average, all of our cascade variants demonstrate significant cost efficiency. In particular, as shown in the average plot, the four MoT variants achieve comparable task performance (∼0.929 accuracy) to GPT-4-CoT-SC (0.931) while demanding only 40% of its cost.

- Sampling from diverse prompt settings helps cascade decision-making. Mixing thought representations is particularly effective. Intuitively, this is because different thought representations can bring in more diverse "opinions" of the weaker LLM on the same input question, resembling how a group of experts with diverse perspectives could contribute to more effective results in collaborative work.

4 Future work

We identify some potential avenues for future research. One extension could be focusing on extending our methods to broad tasks. We could generalize the answer consistency checking to some more complicated applications and integrate other metrics, such as semantic similarity, to evaluate the consistency of the general textual generation tasks.

Another possible extension could be to generalize our approach in complex reasoning tasks to facility question decomposition. In complex reasoning tasks, it is usually necessary to decompose the question into multiple subquestions. Knowing what granularity of decomposition is sufficient and when to stop the decomposition is vital. We can use the answer consistency with different representations to determine whether LLM can solve those subquestions. If not, further processing may be required, such as continuing decomposition.

References

BIG bench authors. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Lingjiao Chen, Matei Zaharia, and James Zou. 2023. Frugalgpt: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. 2022. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. PAL: Program-aided language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10764–10799. PMLR.

Aitor Lewkowycz, Anders Johan Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Venkatesh Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. Solving quantitative reasoning problems with language models. In *Advances in Neural Information Processing Systems*.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167, Vancouver, Canada. Association for Computational Linguistics.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2023. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. In *International Conference on Learning Representations (ICLR)*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Li Zhang, Hainiu Xu, Yue Yang, Shuyan Zhou, Weiqiu You, Manni Arora, and Chris Callison-burch. 2023. Causal reasoning of entities and events in procedural texts. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 415–431, Dubrovnik, Croatia. Association for Computational Linguistics.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*.

A Experiment

A.1 Experimental Setting

We evaluate our LLM cascade approaches on six datasets, covering (1) mathematical reasoning, including GSM8k (Cobbe et al., 2021), ASDIV (Ling et al., 2017), and TabMWP (Lu et al., 2023); (2) symbolic reasoning from BIG-Bench Hard (bench authors, 2023), including DATE and Navigate; and (3) causal reasoning, including CREPE (Zhang et al., 2023). In our pipeline, we leverage the GPT-3.5-turbo (4k context) as the weaker LLM and the GPT-4 (8k context) with CoT self-consistency (Wang et al., 2023, SC) as the stronger LLM. Throughout our experiments, we set the number of task demonstrations as $M = 8$. We set the number of sampling paths as $K = 20$ for GPT-3.5-turbo and $K = 3$ for GPT-4. The sampling temperature by default is 0.4 for both LLMs. The metrics we use are the task accuracy and the relative cost compared with the cost of GPT-4 with CoT SC (denoted as `GPT-4-CoT-SC`).

A.2 Main Results

Figure 3 illustrates the performance of our proposed approaches. For `Vote`-based approaches, we draw curves by changing the pre-defined threshold

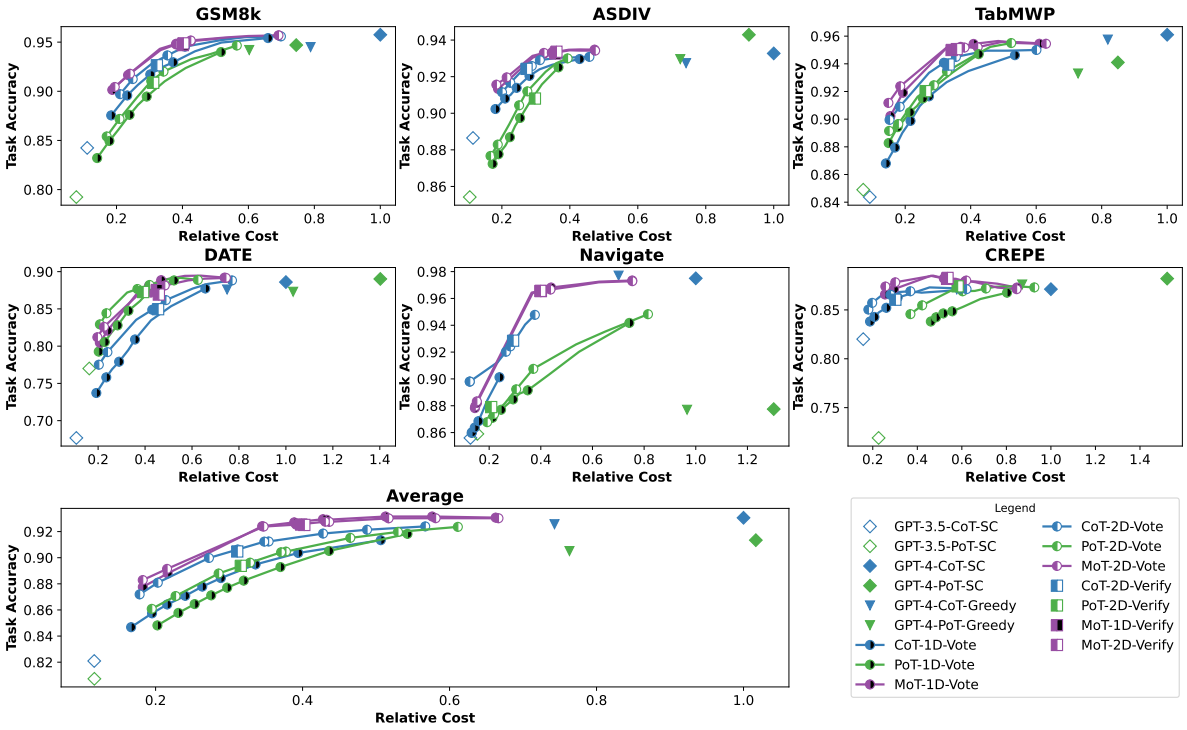

Figure 3: Main experiment results over six reasoning datasets. The bottom figure represents the average performance.

τ varying from 0.4 to 1. A high value of threshold signifies a more rigorous criterion for trusting the answers from the weaker LLM, making more examples transferred to the stronger LLM. Our observations are as follows:

Our pipeline achieves comparable task performance with significantly reduced costs. On average, all of our cascade variants (`Vote` or `Verify`) demonstrate significant cost efficiency. In particular, as shown in the average plot, the four `MoT` variants achieve comparable task performance ($\sim$0.929 accuracy) to `GPT-4-CoT-SC` (0.931) while demanding only 40% of its cost.

Sampling from diverse prompt settings helps cascade decision-making. Our results show that variants involving diverse sources of sampling, such as `CoT/PoT-2D-Vote` and `MoT-1D/2D-Vote`, can more precisely distinguish between easy and hard reasoning questions, compared with their counterparts sampling from single sources, i.e., `CoT/PoT-1D-Vote`. For example, between `CoT-2D-Vote` and `CoT-1D-Vote`, the former outperforms the latter by 1.4% absolute accuracy under the same relative cost of 0.4 on average.

Mixing thought representations is particularly effective. Furthermore, we find that mixing the two thought representations (i.e., `MoT-1D/2D-Vote`) outperforms decision-making using either of them (i.e., `CoT-1D/2D-vote` and `PoT-1D/2D-vote`). Intuitively, this is because different thought representations can bring in more diverse "opinions" of the weaker LLM on the same input question, resembling how a group of experts with diverse perspectives could contribute to more effective results in collaborative work. We also note that when using `MoT`, no obvious difference is perceived between using one set (i.e., `MoT-1D-Vote`) or two sets (i.e., `MoT-2D-Vote`) of task demonstrations. This result reveals that tuning the thought representations is more helpful for measuring an LLM's (un)certainty on its answer than tuning the task demonstrations.