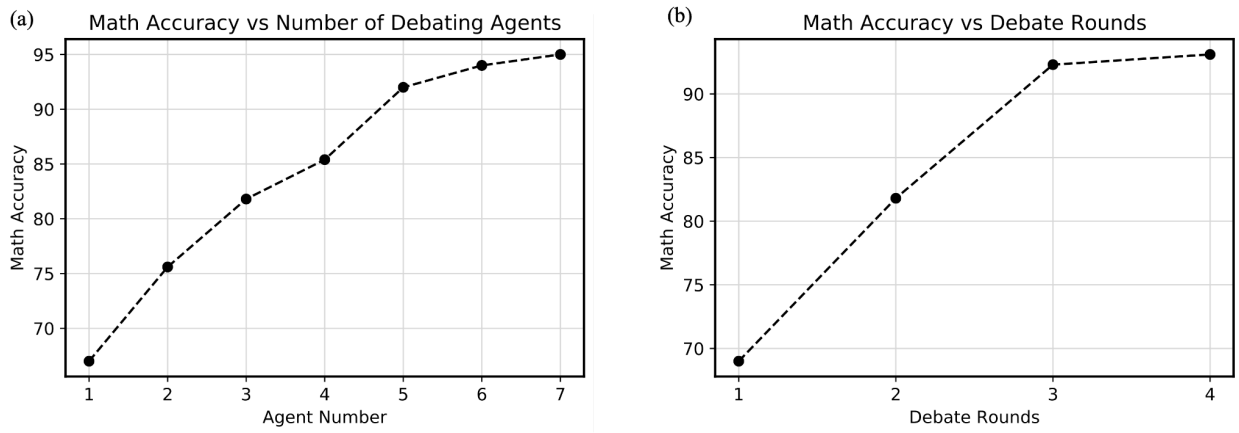
Example Question: Language Models

Improving Factuality and Reasoning in Language Models through Multiagent Debate

In this problem, your task is to read this paper about improving the reasoning abilities of LMs by using multiple agents [1] and getting them to "debate" with each other. We want you to give us a brief summary of this paper, followed by some of your thoughts and takeaway about it. You are free to use whatever open-source framework for this part.



Results from [1] using GSM8K problem dataset.

Using any open-source model of your choice (please confine your choice of models to those that can be obtained from HuggingFace), try to construct the same or, if you want, a *similar* experiment (you can be a bit creative), where you demonstrate increasing accuracy as you scale up the number of agents and debate rounds in a multi-agent debate setting. You are not expected to match the rough numbers reported in the paper, which used commercial models, but it would be impressive if you can show similar performance using only open-source models.

In this exercise, we also want to evaluate how you design and conduct your experiment (especially if you use a different setup than the paper), and report your results in an engaging way as if you are a researcher or engineer writing up a draft section of a part of a short technical report. We want to evaluate your writing and presentation skills too.

References

[1] Preprint: https://arxiv.org/abs/2305.14325,
ICLR 2024 submission (might be more updated than preprint):
https://openreview.net/forum?id=QAwaaLJNCk
ICML 2024 camera-ready version: https://openreview.net/pdf?id=zj7YuTE4t8
Project website: https://composable-models.github.io/llm_debate/