Paper Replication Anthropic - Encouraging divergent thinking in LLMs via multi-agent debate

Hari Narayan

1 Experiment Replication with Anthropic Results

1.1 CounterIntuitive Arithmetic Reasoning Task

Prompting Strategy	GPT-3.5-Turbo	Claude- 2.1	Claude-3-Opus
Zero-Shot(IO Prompt)	20%	22%	**66%**
Zero-shot SC (k=5)	30%	36%	62%
CoT Prompt	24%	34%	50%
Self-Reflect	20%	24%	48%
MAD	36%	26%	48%

Table 1: **Accuracy Scores** for Counter-intuitive Arithmetic Reasoning Task: GPT-3.5 results are adjusted from the paper (not replicated). **Accuracy reported for Zero-shot GPT-4 is 52%**. Score 1 for correct solution else 0.

.

- Anthropic Claude-3.1 Zero shot(66%) gives better results than GPT-4 Zero shot (52%).
- CoT prompt leads to a performance deterioration 66% to 50%. Observed that model inherently replies in CoT style [including Claude-2.1]

  Prompt: *Answer this question by breaking down the question into manageable parts, making necessary assumptions and thinking step by step. Repeat the final answer in -answer–/answer- XML tags.*
- Self-Reflect/ MAD leads to accuracy drop [66% to 48%]. Related read: [1]. This paper shows that SoTA-LLMs have a high tendency to flip their answer when asked "Are you sure?".

2 Remarks

1. **What is the difference between this paper and the other multi-agent debate paper** *Improving Factuality and Reasoning in Language Models through Multiagent Debate* (IFRLM) [2]?

   (a) In the context of a debate, how are the opposing viewpoints generated?

   MAD: The initial solution proposed is being preserved in the chat history, shared to the opponent and the opponent is forced to disagree with the initial solution.

   Prompt: (initial-answer).You disagree with my answer. Provide your answer and reasons.

   IFRLM: The opposing solutions are independently generated.

2. Questionable tricky samples in the Counterintuitive AR dataset:

   - Consider problem No: 3 (0 based index-2) in the dataset:

     Problem: One peach costs one cent. You can use 3 peach pits to exchange for one peach. If you have 10 cents, then what is the maximal number of peaches you can eat?

     Answer: 15

     Explanation: With 10 cents, you can initially buy 10 peaches. After eating these, you'll have 10 pits. You can exchange 9 of these pits for 3 more peaches. After eating these, you'll have 4 pits left (1 from before and 3 new ones). You can then exchange these for another peach. Now, you can eat $10 + 3 + 1 = 14$ peaches. After eating these, you have 2 pits left (1 from before and 1 new pit). Because we are considering the number of peaches that are eaten the most. **You can borrow one pit from others**, and then exchange all the 3 pits for 1 more peach. After eating the last peach, you return the last pit. So in total, you can eat $10 + 3 + 1 + 1 = 15$ peaches.

     Concern: If borrowing is in the picture, whats stopping us from borrowing more? Why stop at borrowing 1? Personally, I feel that 14 is a valid answer.

     What would you expect from a human as response?

   - Consider problem No: 25 - How many times can you subtract 10 from 100?

     Answer should be 1 or 10?

References

[1] Philippe Laban, Lidiya Murakhovs'ka, Caiming Xiong, and Chien-Sheng Wu. Are you sure? challenging llms leads to performance drops in the flipflop experiment, 2024.

[2] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate, 2024.