Paper Replication Anthropic - ReAct

Hari Narayan

1 Experiment Replication with Anthropic Results

While the paper uses Palm-540B a pre-trained LLM model not fine-tuned for chat, the Anthropic Cluade-3-Opus is a chat-LLM model.

1.1 FEVER

Prompt Strategy \Model	Palm-540B	Claude-3-Opus
Zero-Shot(IO Prompt)	57.1%	64%
CoT	56.3%	62.6%
CoT-SC	60.4%	-% %
Act	58.9% %	63.6%
Re-Act	60.9%	63.6%
CoT-SC - ReAct	64.6%	%
Re-Act - CoT-SC	62.0%	%

Table 1: **Accuracy Scores** for FEVER Dataset: Palm-540B results are taken from the paper (not replicated).

.

- **Low recall** of automated LLM-response evaluation is a source of uncertainty regarding the efficacy of prompting technique. Worth studying the actual cause of the errors and how to perform better evaluation. Ideally, the best way would be for a human to perform the evaluation [tedious, time-consuming, not scalable].
    1. *For example, the accuracy of ReAct with the evaluation code as per the paper was 55.1%*
    2. An altered evaluation code snippet **increased the accuracy** by **8 %** by increasing recall to the current reported value of 63.6%.
    3. **Hence the need for robust evaluation of LLM responses to get clear idea about the model capabilities.**
    4. SOTA models as evaluators might be cost-wise prohibitive.

5. **Can an Agentic system based of Open Source Model [2B model-Google Gemma] run on local machine/ Google Colab** be trusted with such evaluations?
6. Can it mix automatic eval procedures currently employed along with natural language understanding for better evaluation?

- Few questions regarding Brazzers, a company in porn industry raised ethical concern related response denial in ReAct(other extended compute promptings as well-Check) but did not rise in standard, zero-shot setting [relate Jail-break].

1.2 HotpotQA

Prompt Strategy \Model	Palm-540B	Claude-3-Opus
Zero-Shot(IO Prompt)	28.7%	33.6%
CoT	29.4%	37 %
CoT-SC	33.4%	- %
Act	25.7% %	39%
Re-Act	27.4%	37.8 %

Table 2: **Accuracy-Exact Match** for Hotpot-QA dataset: Palm-540B results are taken from the paper (not replicated).

.

- Act performs better than or statistically close to Re-Act. Also reported by others: Official Re-Act Github Repo Issues
- Low recall of automated LLM-response evaluation brewing uncertainties. Worth studying the actual cause of the errors.
- Re-Act alone performs worse than Zero-shot for Palm-540B [as reported by the paper] but not observed for Claude-3-Opus.

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) [1]?
   (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] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate, 2024.