Methodology Writeup

Severin Bratus, David Dinucu-Jianu, Ilias Mc Auliffe

We added the following contributions on top of the existing open source OpenDeepResearch framework.

CodeAgent. After switching from the ToolCallingAgent to the CodeAgent we found that the agent did much better on arithmetic tasks and in overall accuracy.

Self-consistency/ensemble. We query the OpenDeepSeek agent multiple times (in the end we chose 12) to mimic an ensemble of agents and then finally combined their predictions using a “Judge” agent which took the query, the context and the responses of each agent in the ensemble and composed the final answer. To maintain computational efficiency, we stop generating agents in the ensemble if the same answer has been provided 4 times in a row. This means that the upper limit of 12 agents is rarely used for only the most difficult queries and most queries only need 4 queries. This further provides an indirect measure of uncertainty of the answer. This had a big impact on accuracy, and we noticed that on subjectively easier queries there was more agreement on the answers across the agents of the ensemble while for more convoluted questions there was disagreement.

Change base LLM. We experimented quite a lot with this. We tried qwq-32b, qwen2.5 72b, llama-v3p3-70b-instruct and qwen2p5-72b-instruct. We found that while all did relatively well, qwq-32b as the CodeAgent and the qwen2p5-72b-instruct as the Judge (since it didn’t require advanced reasoning) were the best combination achieving an accuracy of 71% on the dataset together with the previous two changes (evaluated with gemini-flash).