Report: Sam's work trial test
Basil Labib

Title of work:
Building a custom LLM-based agent to test LLMs on a subset of the MMLU dataset

Report Summary:
In this project, we built a custom agent class which can randomly sample 'n' questions from a subset of the MMLU dataset and prompt a test model for its response. It then compares the model output with the correct answer and keeps track of the performance which it reports at the end of the experiment. The seed model (the model coupled with the scaffolding within the agent) used is **meta/meta-llama-3-8b-instruct** and the test models used were:

1. meta/meta-llama-3-8b-instruct
2. meta/llama-2-7b-chat
3. mistralai/mistral-7b-instruct-v0.2

**OpenAI models could not be used due to the request being constantly blocked by the API.*

Outline of Methodology:

- After studying both METR agents and LangChain agent documentation, we decided to bootstrap with a minimal METR agent scaffold which we incrementally built over time.
- Next, we downloaded a small subset of the MMLU dataset from Hugging Face and parsed the parquet datafiles into Pandas DataFrames. This served as the question bank for testing the models.
- After few minor technical difficulties, we were able to create the testing pipeline which included few agent parameters:
    - seed_model: The model id to be used as the orchestrator within the agent
    - test_model: The model id to be used for testing
    - n_questions: Number of questions to be asked in this experiment
- The Agent API is as simple to use as the following:

```
def main():
      TEST_MODEL1 = "meta/meta-llama-3-8b-instruct"
      n_questions = 25
      agent1 = Agent(df, test_model=TEST_MODEL1)
      print(agent1.evaluate_model(n_questions))
```

Results
The results of the experiment conducted on the above three models for 45 randomly sampled question is given in the following table:

Model id	Total questions	Correct	Wrong	Errors	Accuracy
meta/meta-llama-3-8b-instruct	45	44	1	0	97.77%
meta/llama-2-7b-chat	45	38	7	0	84.44%
mistralai/mistral-7b-instruct-v0.2	45	41	4	0	91.11%

Remarks: As expected, Llama 3 has a superior performance than the other two models. Moreover, the Mistral model seems to be outperforming the Llama 2 model but this observation should be more thoroughly investigated.

Remarks: The choice of Llama 3 as the agent LLM ("seed model") was taken prior to these results but after basic empirical exploration of the performance of few Llama models.

Hiccups

One of the major hiccups was not being able to access the OpenAI models in spite of multiple tries. The debug information pointed towards a rate limit error even when it was the first time querying the API! One possible reason might be that the key provided was out of quota.

Scope for improvements

1. The dataset used is a subset of the MMLU dataset. This decision was made keeping the scope and timeline of the project in mind. The full dataset may be used for a more comprehensive experiment.
2. The MMLU dataset has MCQs from over 60 categories of questions ranging from philosophy to anatomy. A more comprehensive study would be to see the test model's performances category wise.
3. The number of questions was limited to 45 due to limited resources and time. We may check for accuracy over a larger sample.
4. The current agent pipeline works by defining a protocol between the evaluation code and the seed LLM which is used to condition the LLM before the experiment. A future course of action might decouple this step even further to add more flexibility to the pipeline and allow the evaluator to "program" the agent using natural language.
    a. *For example: The current get_question() function which gets a random question from the dataset corpus relies on a mapping from integer to functions in order to check which function to call.*

Code

The agent is publicly available at this repo.
The running log can be accessed here.