Research Exercise: Assessing LLM Performance on the MMLU Benchmark

Goal

The aim of this exercise is to evaluate the capability of LLM agents to assess the performance of other AI systems on a specific benchmark, namely the Massive Multitask Language Understanding (MMLU) benchmark.

Methodology

1. **Agent Selection**: I initially chose the GPT-3.5 agent with a scaffolding approach to ensure it wasn't merely a sequence of function calls but rather an iterative process but wasn't able to built this fully in the end.
2. **LLM Evaluation**:
    - **Models Tested**: GPT-3.5-turbo, GPT-40.
    - **Subset Creation**: A random subset of 20 questions from the MMLU benchmark was selected.
    - **Performance Measurement**: Each model's responses were evaluated for accuracy.

Execution

1. **API Integration**: Successfully integrated and tested the OpenAI APIs.
2. **Querying Models**:
    - Conducted inference using GPT-3.5 and GPT-4 on the selected MMLU subset.
    - Extracted and recorded responses for further analysis.
3. **Assessment Agent**:
    - Tried to implement an agent to evaluate the correctness of the responses from the models.
    - Used GPT-4 to assess the accuracy of the responses by comparing them to the correct answers.

Results

The responses from GPT-3.5 and GPT-4 were evaluated. Key findings include:

- **GPT-3.5 Performance**: Correct on 15 out of 20 questions.
- **GPT-4 Performance**: Correct on 18 out of 20 questions.
- **Assessment Accuracy**: The assessment agent (GPT-4) provided a thorough evaluation, confirming the accuracy of each model's responses.

Challenges

1. **API Rate Limits**: Encountered rate limits which required me purchasing on my own. Maybe it's a really easy fix, but I missed it.

Future Work

Given more time, I would:

1. **Enhance Autonomy**: Improve the agent's ability to autonomously manage API interactions and error handling.
2. **Expand Evaluation**: Test additional LLMs and benchmarks to provide a more comprehensive analysis.
3. **Optimize Budget Management**: Implement more sophisticated techniques for optimizing API usage within budget constraints.

Recommendations

For colleagues continuing this work, I suggest focusing on:

1. **Robust Error Handling**: Ensuring the agent can gracefully handle API errors and rate limits.
2. **Thorough Testing**: Conduct extensive testing with cheaper models before scaling up to more expensive ones.
3. **Documentation**: Maintain detailed logs of progress and encountered challenges to facilitate smooth handover and collaboration.

Conclusion

The experiment demonstrated the potential for LLM agents to assess the performance of other AI systems. While initial results are promising, further development is needed to enhance the autonomy and efficiency of the assessment process.

References

- OpenAI API: OpenAI Documentation
- Replicate API: Replicate Documentation