Sprint 1 Report (10/05/2024 - 11/02/2024)

What's New (User Facing)

- Initial framework setup for LLM evaluation tool.
- Foundation for inserting new tests and adding LLMs.
- Research on existing LLM performance metrics and testing protocols.

Work Summary (Developer Facing)

During Sprint 1, our team focused on laying the groundwork for the LLM evaluation tool by conducting research and developing the base functionality of our Python framework. We gathered client requirements, created UML diagrams, and established user stories to guide future development. We also installed and evaluated existing solutions to better understand how they could inform our design. The team faced some barriers due to limited access to paid materials, which hindered our ability to access full learning resources, but we worked around it by using publicly available alternatives. Overall, this sprint established a solid foundation for further development.

Unfinished Work

We were not able to complete the integration of the evaluation metrics within the framework as planned. This was due to unexpected complexities in the metrics research and our team's limited access to some required materials. These tasks will be moved to the next sprint.

- Progress on evaluation metrics tracked but not fully implemented.
- Commented on the issues regarding limited resource access.

Completed Issues/User Stories

Here are links to the issues that we completed in this sprint:

- Create Standardized Querying Function
- Set Up Python Environment
- Evaluation and Improvement of LLM Accuracy Using Insights from DeepLearning.AI Course
- Automated Evaluation of LLMs for Improved Performance
- Researching The Mathematics Behind LLMs
- LLM Catastrophic Forgetting Guest Lecture
- LLM Coursera Background Research Week 1
- LLM Coursera Background Research Week 2
- Write Initial Project Documentation
- Research Factual Accuracy Metric
- Compile a list of the major LLMs
- Research API Documentation

Incomplete Issues/User Stories

Here are links to issues we worked on but did not complete in this sprint:

- Implement Logging Mechanism — The logging implementation was delayed due to critical bug fixes and ensuring compatibility with our centralized monitoring systems.
- Implement Secure API Authentication - The secure authentication implementation was delayed due to ongoing work on securely storing API credentials and ensuring compatibility with various authentication methods across different LLMs.
- Enhance LLMEval Class with API Parameter Control - The enhancement of the LLMEval class was delayed due to ongoing testing of parameter compatibility and ensuring backward compatibility with existing API calls.
- Research Ethical Considerations Metric - The investigation into ethical evaluation techniques was delayed due to prioritizing urgent project deadlines and gathering comprehensive research on existing bias detection tools and toxicity models.
- Research Relevance Metric - The research into relevance metrics was delayed due to focusing on other critical tasks and the need to review extensive NLP literature and tools for effective implementation.
- LLM Coursera Background Research Week 3 - The creation of standardized testing protocols was delayed due to prioritizing the completion of relevant training materials and ensuring a thorough understanding of reinforcement learning from human feedback principles.

Code Files for Review

Please review the following code files, which were actively developed during this sprint, for quality:

- Models.py
- LLMEval.py
- Test_LLMEval.py
- Test_model.py

Retrospective Summary

Here's what went well:

- Clear client communication helped us establish firm project requirements.
- The base framework for the tool was completed successfully.
- Team members adapted well to researching performance metrics and learning new skills.

Here's what we'd like to improve:

- Access to more comprehensive learning resources, either by seeking alternatives or clarifying free options with the client.
- Better time estimation for tasks involving research or complex integrations.

Here are changes we plan to implement in the next sprint:

- Begin the integration of semantic analysis, BERT, and BLEU metrics into the framework.
- Complete research into remaining evaluation techniques and refine user stories for further development.