Sprint 2 Report (10/6/24 - 11/5/24)

What's New (User Facing)

- Perplexity Evaluation Score
- Bleu Evaluation Score
- Rouge Evaluation Score
- Meteor Evaluation Score
- chrF Evaluation Score
- Wer Evaluation Score
- Edit distance Evaluation Score
- BertScore Evaluation Score
- Word Movers Distance Evaluation Score
- Universal Sentence Encoder Similarity Evaluation Score
- Fact score Evaluation Score
- Type-Token Ratio (TTR) Evaluation Score
- MTLD (Measure of Textual Lexical Diversity) Score
- HD-D Evaluation Score

Work Summary (Developer Facing)

Our primary focus this sprint was building evaluators for several of the evaluation metrics requested by the client. To complete this task, we had to do some more research on which metrics are commonly used in industry and academia, and how we can implement them. We completed this research and selected the evaluators that are most aligned with the client's needs. The completed evaluators are listed in the What's New section. All of these evaluators take in LLM data and generate a quantitative score for one or more of the evaluation modules. Preliminary research has been completed on the other evaluators, but further research and discussion with the client may be necessary before implementation. We also completed the backend for querying LLMs automatically and receiving the data. The Evaluation Handler will allow us to run different evaluators simultaneously with the LLM-generated data using multi-threading; this functionality is partially implemented.

Unfinished Work

This sprint we were unable to complete the following evaluators: Cosine Similarity in Embedding Space, Universal Sentence Encoder Similarity Metric, N-gram Overlap, Grammater Metric(grammarly api), Relevance Scores, Topicality and Focus, Mean Opinion Score, Dialog Turn-Level Metrics, Knowledge F1 Score, Factuality Evaluation for QA, TruthfulQA, Attributable Factuality, Toxicity Detection, Bias and Fairness Metrics, Unintended Bias Measures,

Offensiveness Score, Safety Checklist Compliance, Adversarial Testing, Perturbation Sensitivity, Error Sensitivity, Mean Length of T-Unit, Clauses Per T-Unit, Depth of Syntactic Tree, Syntactic Diversity, Latent Semantic Analysis, Word Embeddings Novelty, Topic Modeling, Semantic Surprise, Divergent Association Testing, Remote Association Testing, Alternative Uses Testing, and possibly more. We were unable to complete these due to running out of time, for as you can see there are a significant number of evaluators that can be created, yet only so much time in a sprint. Additional unfinished work is inside of the system; the evaluator handler is only partially built as the construction of it began after several evaluators were created in order to develop the optimal format for its construction. Then the backend storage for evaluation metrics is not yet built as that is dependent on the handlers responses. All of this work is necessary, so will be added to the next sprint.

Completed Issues/User Stories

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

- https://github.com/RyderSwanson/LLMEval/issues/38
- https://github.com/RyderSwanson/LLMEval/issues/39
- https://github.com/RyderSwanson/LLMEval/issues/40
- https://github.com/RyderSwanson/LLMEval/issues/41
- https://github.com/RyderSwanson/LLMEval/issues/42
- https://github.com/RyderSwanson/LLMEval/issues/43
- https://github.com/RyderSwanson/LLMEval/issues/44
- https://github.com/RyderSwanson/LLMEval/issues/45
- https://github.com/RyderSwanson/LLMEval/issues/46
- https://github.com/RyderSwanson/LLMEval/issues/47
- https://github.com/RyderSwanson/LLMEval/issues/49
- https://github.com/RyderSwanson/LLMEval/issues/50
- https://github.com/RyderSwanson/LLMEval/issues/51
- https://github.com/RyderSwanson/LLMEval/issues/14
- https://github.com/RyderSwanson/LLMEval/issues/26
- https://github.com/RyderSwanson/LLMEval/issues/28
- https://github.com/RyderSwanson/LLMEval/issues/29
- https://github.com/RyderSwanson/LLMEval/issues/27

Reminders (Remove this section when you save the file):

- Each issue should be assigned to a milestone
- Each completed issue should be assigned to a pull request
- Each completed pull request should include a link to a "Before and After" video
- All team members who contributed to the issue should be assigned to it on GitHub
- Each issue should be assigned story points using a label
- Story points contribution of each team member should be indicated in a comment

Incomplete Issues/User Stories

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

- https://github.com/RyderSwanson/LLMEval/issues/48
- https://github.com/users/RyderSwanson/projects/2?pane=issue&itemId=86149785&issue=RyderSwanson%7CLLMEval%7C52

- We did not get to these issues because further research was required before implementing different evaluators.
- We also had limited time this sprint so we prioritized the evaluators that could be implemented relatively easily and demoed for the client at a later date.

Code Files for Review

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

- EvaluationModules.py
- EvaluationHandler.py
- demo.py
- evaluatorDependencies
- LLMEval.py

Retrospective Summary

Here's what went well:

- Research of various evaluation metrics.
- Implementation of various evaluation metrics.
- Meetings and communication with client.
- Backend work for querying LLMs and managing data.

Here's what we'd like to improve:

- More testing for evaluation metrics.
- More documentation for evaluation metrics.

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

- Add testing and documentation alongside metric implementation (rather than after)

- Separating the evaluators into different modules (e..g creativity, relevance, coherence, etc.)
- Testing different evaluators as they are implemented.