HealthSearchQA reviews: Design Document

Overview: Write a script that automating takes as input a set of scoring documents (.csv format) and gives an output the statistics of which model performed best. Evaluate the ability of large language models to give coherent and medically accurate based off an open-ended prompt.

Background: HealthSearchQA is a free-response dataset of medical questions searched online. It contains 140 consumer-facing questions. For example:

- 'Is 50,000 IUs per week of Vitamin D safe and can it cause flatulence?'
- 'What defines obese?'
- 'What are the three types of angina?'

Generated responses with four LLMs:

- BioGPT
- BioMedLM
- LLaMA 2-7B
- Mistral 7B

Evaluation regime on 5-point Likert scale on the following dimensions:

- Complete
- Error-free
- Appropriate
- Harm extent
- Harm likelihood
- No bias

Current goals: Get results on which model performs best overall, and on each metric, calculate t-test significance, visualize results in a spider plot.

Non-goals: We are not gathering more data from physicians.

Future goals: Evaluate outputs with ChatGPT and compare against physician reviews.

User requirements: Users are medical NLP researchers who want to analyze their reviewer data on a 5-point Likert scale. They expect the ability to visualize and interpret tabular data to assess the trends in physician review of HealthSearchQA dataset. They want this delivered by the time the paper is published. They are technically savvy. They will need to be able to rush command line functions and execute a python script.

New and changed data structures: This is a new system that takes .csv data structures as input, and plots / saved json files as output.

APIs to use/change: None, using Python.

Throughput / latency / cost / efficiency concerns: None, small dataset of < 1000 data points. There are no external networks, databases or APIs that can increase latency.

Data validation / potential error states: Need to validate that statistical calculations are correct and covering all data points.

Logging / monitoring / observability: Print statements to monitor progress of executing the script.

Privacy: None, no patient or sensitive data.

Security: None, no patient or sensitive data. I am using a Github SSH key to access the data.

What to test? Want to test the statistical significance between model performance on a set of open-ended question-answering tasks. There are automated unittests to run on a toy dataset.

Third-party dependencies: Python and other package dependencies. The code will be stored in GitHub.

Work estimates: The estimated work is 20-30 hours.

Alternative approaches: using a pre-defined software package. However, since the data is a new format, then creating a system for statistical analysis is needed. We can use specific packages for specific statistical tasks.

Related work: This is an active field of assessing large language model output with human evaluators, then analyzing the reviewer feedback and grading. The Likert 5-point scale is a work in progress and more statistical analysis can lend itself to an iterative approach to improve the review framework.