Project Proposal: Measuring LLM Responses

Anonymous ACL submission

Abstract

What are the attributes of text generated by large language models, and which ones make them useful?

1 Introduction

Speed can be measured in terms of time to produce a complete response or in terms of time per token.

2 Literature Review

Speed:

https://arxiv.org/pdf/2305.15038.pdf provides some speed benchmarks on data analysis tasks in Table 4 on Page 6. It doesn't really discuss methodology, though.

Context window

3 Data

For some questions, new data must be collected.

Annotation: Accuracy For missed/incorrect answers, location of the error

For other questions, pre-existing sources of LLM prompts and responses can be used.

Scope of Data:

Going for the same broad set of fields in the original GPT-4 technical paper. https://arxiv.org/pdf/2303.08774.pdf

Academic - General (not Subject-Specific) Standardized tests (All GRE Sections) Coding - Leetcode Easy, Medium, Hard Other Common NLP Benchmarks

Models:

GPT-4 Claude Bard

Consider the extent of repetition of the same prompt to the same model.

4 Responsiblities

I am the only author of this proposal, but I am open to working with others with similar interests.

Limitations

The actual length of time required by an online LLM to answer a question is difficult to estimate given the presence of confounding factors such as connection speed and server load. To limit associated variation, responses were generated during off-peak hours.

Acknowledgements

The template for this document was adapted by Jordan Boyd-Graber, Naoaki Okazaki, and Anna Rogers.