Project in 02445:
Design and Implement a Evaluation Scheme for Large Language Models

June 11, 2024

This project involves developing a systematic evaluation scheme for large language models (LLMs) to assess a desired facet of an LLM for instance, but not limited to, fluency, coherence, contextual relevance, consistency, bias, robustness. The goal of this project to gain practical experience in evaluating real-world AI systems and put in practice the statistical and and other topics touched upon in the course.

1 Resources

- Evaluating gender-bias in STEM suggestions from LLMs - shared as

  `LLM_bias_1_revision.pdf`

  In this work, ChatGPT is tested for bias without any ground-truth data. The (qualitative) data is generated by repeated sampling using 4 different prompts.

- Evaluating large language models: a comprehensive survey - shared as

  `LLM_evalaution_survey.pdf`

  This serves as a comprehensive guide to LLM aspects, metrics, datasets etc.

2 Format of the report

1. Maximum 10 pages. Ideal 6-7 pages.
2. Include absolutely improtant information in the main body. Details that you think may be important but are unsure can be put in the appendix. I am not obliged to look at the appendix.
3. If code, put it in the appendix.
4. For fairness in page considerations, use the report template shared:

   `02445_report_template`

3 Steps to consider while designing your evaluation

1. Which aspect of the LLM are you going to evaluate?
   (a) What metric will you use? - Please describe the aspect and metric sufficiently in your report
2. Which LLM did you choose on?
   There are large number of options to choose from, both online and offline. ChatGPT and derivates, Bing Copilot, etc. These are examples of online systems, therefore when using these (or any online system), please consider the following points:

- Is there a daily/overall limit in the number of prompts you can make in the non-paid versions?
- Remember, most of the online LLM based tools are continuously learning from your prompt. How can you ensure your samples are independent? One approach could be to set your privacy settings not use your data.

There are multiple offline API as well, that can be locally run. If you choose to work with local/offline systems, please consider the runtime resources (eg: does your laptop need to have a GPU, etc) required.

3. Data for evaluation

   - Are you using an available benchmark dataset?
     Describe the data, features attributes in the report.
   - Are you collecting data yourself (as in the share resource on LLM bias)?
     (a) Describe the design of your prompt.
     (b) Are you including any additional factors of variation (Eg: subgroups like age, gender, domain, etc)? Why?
     (c) How will you convert the LLM responses to data that you can model using a probability distribution?
     (d) How are you deciding on the number of times you will prompt the LLM? (hint:sample size estimation, if applicable.)

4. Results

   (a) Discuss the qualitative aspect of the results
   Wherever applicable, feel free to present plots (barplots, boxplots, etc) and discuss your observation of the outcome.

   (b) Quantitative analysis
   Can you statistically quantify and compare the metrics you are testing the LLM for? Sufficiently describe why the statistical test is relevant, what the assumptions for the test are, are they completely/approximately met.