CS510 Project Proposal

Project Track: Research Track
Team Members:

zhent6@illinois.edu Zhen Tang [Coordinator/lead]
qilong3@illinois.edu Qilong Wu
xhuang8@illinois.edu Xiao Zhuang
kevinez2@illinois.edu Kevin Zhang

[Research Question] What is your research question? Clearly define the research problem/question.

Motivation: With AI-generated content becoming more and more sophisticated and human-like, traditional means of training models to detect AI-generated content are becoming less and less feasible. Because of this, the concept of AI watermarks has been introduced to identify AI-generated content. Prior work has been done to test how watermarking affects an AI's output, specifically in the context of random generation of values. However, the baseline random generation of values is still unexplored.

Primary research question: Can we use large language models to reliably generate random data?

Sub Questions: If not, what are the discrepancies between the results? What are the distributions of the data that it generates? Can prompting affect the outcome of the results?

- [Significance] Why is this an interesting question to ask and why would we care about the answer to this question or a solution to the problem?

  The distribution of the random generation when LLM is prompted, is still underexplored. Also, it is essential to the security of the LLM model, like watermarking. For example, there is ongoing work utilizing random generation to evade watermarking detection for a previously believed secure scheme and its variant [1].

- [Novelty] Has any existing research work tried to answer the same or a similar question, and if so, what is still unknown? (In other words, what is the novelty of your research question?) Provide a brief list of related work.

  One recent work found the distribution when LLM is prompted to do random generation is far from the prompted uniform and model-specific [2]. It is unknown if there are principles for such diverse behavior, e.g. They share similar biases as humans. If we have time, we could potentially come up with a novel evaluation metric for such work.

- [Approach] How do you plan to work out the answer to the question? (At the proposal stage, you are only expected to have a sketch of your methods.)

  We are going to try different prompting strategies on several LLM models to see if it affects the generation of random content. For example, we will give it a list of words or numbers and ask it to pick at random a word from the list. We will conduct this experiment on the same list of words 500 times and investigate the distribution of the words generated to see if there is a bias in the data. Another experiment we can do is investigate how LLMs complete sentences given some sort of prefix. In this case, we will give it a list of words, and a sentence prefix, and ask it to complete the sentence using the words in the list. Again, we will investigate the distribution of the w

- [Evaluation] How would you evaluate your solution? That is, how do you plan to demonstrate that your solution/answer is good or reasonable?

  There are a multitude of ways we can evaluate our results. We can evaluate the word distributions generated by models using current metrics. We can also try to find some features of high-frequency words the model picks.

- [Timeline] A rough timeline to show when you expect to finish what. List a couple of milestones if possible (they can be tentative).
  1. Choose models to conduct experiments on.
  2. Generate data from those models.
  3. Develop evaluation metrics.
  4. Evaluate the results based on those evaluation metrics.
- [Task division] Use one sentence to describe what each team member is expected to work on (can be tentative).

  Everyone will collaborate on all the parts of this project. Each team member will be responsible for choosing a model and conducting the experiments on the model they choose. Everyone will review other members' work and discuss the tough questions together. The group will meet. once a week to discuss the project.

References

[1] Wu, Qilong, and Varun Chandrasekaran. "Bypassing LLM Watermarks with Color-Aware Substitutions." *arXiv preprint arXiv:2403.14719* (2024).

[2] Tang, Leonard, Gavin Uberti, and Tom Shlomi. "Baselines for Identifying Watermarked Large Language Models." *arXiv preprint arXiv:2305.18456* (2023).