Deep Learning Fall 2023 Project Proposal

Seeger Zou
jz3714@nyu.edu

Justin Hwang
jh8032@nyu.edu

Orion Yang
ry2091@nyu.edu

November 16, 2023

1 Problem Statement

Large Language Model proliferation has steered the Artificial Intelligence industry over the past few years. One area of focus on LLMs is their ability to act as roleplay agents and perform tasks specified by their role. As social actors, these models can provide assistance in a wide range of fields. We want to investigate the capabilities of the state-of-the-art model LLaMA 2 [1] as a RoleLLM, and if possible, compare it to research done on the same task using other models. We also plan to address the open-ended nature of the outputs of generative AI models, as they produce more qualitative results; it makes the output hard to evaluate, compared to other predictive ML models that have clear performance metrics. Furthermore, we wish to comment on the ethical issues that arise with commercializing these models as replacements for artists, as seen in recent strikes and calls for regulations.

2 Literature Survey

Recent studies have shown that LLMs can be manipulated to have emotional intelligence [2]. This allows developers to use methods such as instruction tuning to enhance and personalize the model's performance on a given role. While open-source models still lag behind models such as GPT, they show promise in their task-learning abilities. There has been a study published on RoleLLM [3] and to our knowledge, they are among the first to tackle LLM's capability in roleplaying; they proposed the first benchmark for roleplaying models. From their experiments involving multiple agents on multiple tasks, we hope that by narrowing down on one single task, we can add to their findings and come up with new benchmarks.

3 Datasets

Our dataset strategy involves extracting both voice and script from Professor Gustavo's Zoom lecture recordings. This process captures the nuances of his speaking style by isolating audio segments featuring his voice and simultaneously transcribing the spoken content. For specific questions invoking the knowledge base, we plan to extract important information from the textbooks via glossary or online datasets that can be easily incorporated. The transcribed dataset from the recordings serves as comprehensive training data for fine-tuning the LLaMA 2 model [1]. This approach allows the model to learn linguistic nuances, intonation, and pacing, enhancing its ability to generate authentic responses. Furthermore, the audio dataset supports our extended goal of generating answers to queries, and eventually voiced speech, by providing a holistic representation of Professor Gustavo's communication style. By leveraging both voice and script components, we aim to create a robust dataset that enables the model to faithfully replicate his unique style across both textual and auditory dimensions.

4 Models

For our primary experimentation, we will be using the open-source model, LLaMA 2 [1]. Since the model is trained on a diverse dataset, encompassing a broad range of topics and styles, it can understand and generate text across various domains and contexts. Through fine-tuning our specific dataset, we hope to generate immersive, character-consistent, and contextually appropriate responses in a roleplaying scenario. We will assess the model's performance using metrics by focusing on the interaction and engagement of the model, such as the ability to answer various types of questions. Additionally, the quality of the answer can be evaluated by whether the generated response sufficiently reflects the character's speaking style and if the response correctly answers the question [3]. Extended experiments would include using a different model such as BERT or GPT and comparing their results.

5 Goal

Our goal is to develop an LLM capable of generating responses in a manner consistent with Professor Gustavo's speaking style. If time permits, we aim to extend our model to generate not only text responses but also speech clips, by using the audio recordings from the available lecture datasets. Moreover, our model will serve as not only a basis for intelligent agents/assistants in the education field but also as a case study for exploring the ethical considerations associated with the commercialization of such language models.

References

[1] Touvron, Hugo and Martin, et al. (2023) Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288

[2] Li, et al. (2023) Emotionprompt: Leveraging psychology for large language models enhancement via emotional stimulus. arXiv:2307.11760

[3] Wang, et al. (2023) RoleLLM: Benchmarking, Eliciting, and Enhancing Role-Playing Abilities of Large Language Models. arXiv:2310.00746