Project Summary

Overview

The astounding progress in generative AI has led to a marketplace burgeoning with large language models (LLMs), each claiming to be better than the rest. Despite being of paramount importance, the evaluation of LLMs has not kept pace with this progress, making it increasingly challenging to select the best model for a desired user application. We still rely on relatively few downstream tasks and benchmarks, which are not entirely blind to the language model, due to training data contamination. Even the most reliable mechanism of comparing generative models of language—relative human judgment—is widely known to be inconsistent, and subjective. In the absence of a reliable system of comparison, most LLM designers single-handedly focus on training data quantity for model improvement, raising questions about the future of the field.

We hypothesize that understanding and quantifying the relative differences between models will need to consider factors at different levels of abstraction, often ignored in standard evaluation. Due to the differences in the design of training—such as data quantities, mixtures and schemes—each model possesses a unique identity, which determines the tokens it can and cannot generate. Our proposed research shows that this fundamental difference can not only result in a model signature, but also provide an effective mechanism to differentiate model capabilities. The generations from different LLMs, under exactly the same conditions, are often too similar for human judgment; our proposed research determines how we can systematically differentiate candidates based on their suitability for human evaluation. Ultimately the differences between models needs to take into consideration not only their utility for end users, but also their safety and trustworthiness. Our proposed research provides a novel comparison for LLMs based on the degree of alignment of models to preferences set by the LLM designers. My proposed research is organized in three thrusts: understanding the relative capabilities of language models in terms of the differences between (i) what the models can or cannot generate; (ii) human judgments of their generations; and (iii) their alignment with different preference criteria. If successful, all three thrusts will result in improved understanding of the relative capabilities of LLMs.

Keywords: Generative Evaluation; Large Language Models; Language Generation; Interpretability; Preference Alignment; Model Signatures; Human Judgment.

Intellectual Merit

Our proposal evaluates the relative capabilities of LLMs under a more robust framework, providing a cleaner pathway for practitioners to select the right model for their application. These comparisons take into account three different and complementary attributes of LLMs: training and data architecture, human acceptability of their generations and their alignment with provider preferences. Our findings will shed light on three different mechanisms behind LLMs' relative capabilities and limitations. Our research will result in effective metrics and novel evaluation algorithms to enable meaningful and reliable comparisons between very similar LLMs, leading to improvement in user trust and satisfaction, as well as better user safety through alignment with desirable model behavior.

Broader Impacts

Generative models of language are steadily becoming pervasive in our increasingly online world, yet their impact on the society at large remains unknown. This calls for an immediate attention to AI education, where we teach students how AI capabilities differ across models, and why. A key component of our proposal will contribute to the research infrastructure behind model choice: understanding what different models can and cannot do will impact the landscape of applications built on the foundation of generative AI. The PI regularly participates in education efforts aimed at both high school students and teachers, through the USC Viterbi K-12 Center, in collaboration with the Los Angeles County Office of Education. Further, this research will result in more effective, safe and trustworthy AI assistants for a wide variety of tasks designed for a diverse user population. The PI's lab comprises a diverse group of researchers, with members from historically underrepresented groups in computer science.