USC Viterbi School of Engineering
Ming Hsieh Department of Electrical and Computer Engineering

Center on AI Foundations for the Sciences

AIF4S Seminar Series

Language Model Alignment: Theory & Practice

Ahmad Beirami

Google Research

Date: February 16, 2024 (Friday)
Time: 3:30 PM
Location: EEB 132

Abstract:
Generative language models have advanced to a level where they can effectively solve a variety of open-domain tasks with little task specific supervision. However, the generated content from these models may still not satisfy the preference of a human user. The goal of the *alignment* process is to remedy this issue by generating content from an aligned model that improves a reward (e.g., make the generation safer) but does not perturb much from the base model. A simple baseline for this task is best-of-N, where N responses are drawn from the base model, ranked based on a reward, and the highest ranking one is selected. More sophisticated techniques generally solve a KL-regularized reinforcement learning (RL) problem with the goal of maximizing expected reward subject to a KL divergence constraint between the aligned model and the base model. An alignment technique is preferred if its reward-KL tradeoff curve dominates other techniques.

In this talk, we give an overview of language model alignment and give an understanding of known results in this space through simplified examples. We also present a new modular alignment technique, called controlled decoding, which solves the KL-regularized RL problem while keeping the base model frozen through learning a prefix scorer, offering inference-time configurability. Finally, we also shed light on the remarkable performance of best-of-N in terms of achieving competitive or even better reward-KL tradeoffs when compared to state-of-the-art alignment baselines.

Bio:

Ahmad Beirami is a research scientist at Google Research, leading research efforts on building safe, helpful, and scalable generative language models. At Meta AI, he led research to power the next generation of virtual digital assistants with AR/VR capabilities through robust generative language modeling. At Electronic Arts, he led the AI agent research program for automated playtesting of video games and cooperative reinforcement learning. Before moving to industry in 2018, he held a joint postdoctoral fellow position at Harvard & MIT, focused on problems in the intersection of core machine learning and information theory. He is the recipient of the Sigma Xi Best PhD Thesis Award from Georgia Tech.

Host: Dr. Mahdi Soltanolkotbi, soltanol@usc.edu, EEB 430