THE UNIVERSITY OF CHICAGO | DEPARTMENT OF STATISTICS

Master's Thesis Presentation

Peiran Chen
Department of Statistics
The University of Chicago

"Comparative Analysis of LLM Alignment Techniques for Open-Source Model"

May 8, 2025 at 1:30 PM
Jones 111, 5747 S. Ellis Avenue

Abstract

In recent years, Large Language Models(LLMs) have shown promising result and impressive performance on a variety of Natural Language Processing(NLP) tasks. It is also worth noting that it is important to ensure all these models produce outputs in accordance with human values and preferences. Unaligned LLMs can generate information that are toxic, biased, unethical, or factually incorrect, which could leads to serious consequences in real-world applications. To reduce these risks, people have come up with prominent LLM alignment techniques–Reinforcement Learning from Human Feedback(RLHF), Direct Preference Optimization(DPO), and Supervised Fine Tuning(SFT)–aimed at aligning the output of LLMs towards desired outcomes. In this paper, we specifically focusing on their application to open-source LLM Phi-3.5 mini and assessing performance across diverse safety benchmark datasets primarily through refusal rates. To gain deeper insights beyond whether model refuse harmful prompts, we further analyze the misalignment by using TF-IDF and K-Means on the generated responses. Our result shows that while these techniques demonstrably enhance model safety compared to baseline performance, they exhibit distinct strengths and weaknesses depending on the specific safety challenge.

For information about building access for persons with disabilities, please contact Keisha Prowoznik at 773.702-0541 or send an email to kprowoznik@uchicago.edu. If you wish to subscribe to our email list, please visit the following website: https://lists.uchicago.edu/web/info/statseminars.