THE UNIVERSITY OF CHICAGO | DEPARTMENT OF STATISTICS

Master's Thesis Presentation

Zijiang Yang
Department of Statistics
The University of Chicago

"Contrastive Representation Learning for Prompt–Response Alignment"

May 7, 2026, at 4:30 PM
Jones 111, 5747 S. Ellis Avenue

Abstract

Reinforcement learning from human feedback (RLHF) has become a standard approach for aligning large language models with human preferences, but reward models often rely on superficial features such as length, style, or fluency rather than true adherence to prompt intention. This thesis studies prompt–response alignment as a representation learning problem and proposes TextCLIP, a contrastive framework that learns a shared embedding space for prompts and responses. By pulling matched prompt–response pairs together and pushing mismatched pairs apart, TextCLIP aims to capture prompt-specific semantic compatibility instead of absolute response quality. The learned shared embedding space also exhibits meaningful geometric structure, with better-aligned responses tending to lie closer to prompt representations than less compatible ones, a pattern that is consistent with the strong performance of TextCLIP on the on-topic evaluation task.
