THE UNIVERSITY OF CHICAGO | DEPARTMENT OF STATISTICS

Master's Thesis Presentation

Yunsheng Lu
Department of Statistics
The University of Chicago

"Causal Alignment of Reward Models via Response-to-Prompt Prediction"

February 6, 2025, at 4:00 PM
Jones 111, 5747 S. Ellis Avenue

Abstract

Reward models are central to aligning large language models, yet they often overfit to spurious cues such as response length and overly agreeable tone. Most prior work weakens these cues directly by penalizing or controlling specific artifacts, but it does not explicitly encourage the model to ground preferences in the prompt's intent. We learn a decoder that maps a candidate answer to the latent intent embedding of the input. The reconstruction error is used as a signal to regularize the reward model training. We provide theoretical evidence that this signal emphasizes prompt-dependent information while suppressing prompt-independent shortcuts. Incorporating this signal into RM training in Gemma-2-2B-it and Gemma-2-9B-it increases RewardBench accuracy from 83.22% to 86.83%. For Best-of-N selection, our framework increases length-controlled win rates while producing shorter outputs, and remains robust to lengthening and mild off-topic drift in controlled rewrite tests.

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.