THE UNIVERSITY OF CHICAGO | DEPARTMENT OF STATISTICS

Master's Thesis Presentation

Pippa Lin
Department of Statistics
The University of Chicago

"Linear Directions of Political Ideology in LLMs: Probing, Generalization, and Intervention"

May 7, 2026 at 12:00 PM
Jones 111, 5747 S. Ellis Avenue

Abstract

Large language models (LLMs) appear to encode political ideology as a linear structure in activation space. This work examines whether that structure generalizes across three ideology measures: DW-NOMINATE, CFscore, and a Bradley–Terry score. Using Mistral-7B-Instruct, we train linear probes on senator-generated prompts and find that all three scores are recoverable from middle-layer attention heads. The probe directions are strongly related but not identical, suggesting both a shared liberal–conservative axis and score-specific variation. Intervention experiments further show that steering along the DW-NOMINATE and Bradley–Terry directions systematically shifts generated text, while CFscore produces less stable effects. Overall, the results suggest that political ideology in LLMs is structured, approximately linear, and partially dependent on the measurement framework.

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.