THE UNIVERSITY OF CHICAGO | DEPARTMENT OF STATISTICS

Master's Thesis Presentation

Ian Joffe
Department of Statistics
The University of Chicago

"Searching for Linear Representation of Political Sentiment in Large Language Models"

November 17, 2025, at 8:00 AM
Jones 111, 5747 S. Ellis Avenue

Abstract

This thesis searches for a linear representation of political sentiment in the activations of the Gemma 2-2B large language model. To accomplish this, I put together a dataset of congressional bills, and filter it to pick out the especially partisan ones. I attempt to linearly classify activations based on the partisanship of the bill, and to linearly patch the activations with the goal of having the model express support for more left- or right-leaning bills under a role-playing prompt. Ultimately, I find only a small amount of signal even in the filtered dataset, and conclude that more data than I used for classification and patching may be necessary for convincing observations.

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.