BETTER ALIGNED WITH SURVEY RESPONDENTS OR TRAINING DATA? UNVEILING POLITICAL LEANINGS OF LLMS ON U.S. SUPREME COURT CASES

Shanshan Xu[1], Santosh T.Y.S.S[1], Yanai Elazar[2,3], Quirin Vogel[1], Barbara Plank[4], Matthias Grabmair[1]

[1]TU Munich, Germany; [2]Allen Institute for AI; [3] University of Washington;
[4]LMU Munich & Munich Center for Machine Learning (MCML), Germany

Introduction

- LLMs primarily learn from their pretraining data and often memorize its patterns.
- Political bias has been observed in LLM outputs [1]. However, to which extent these biases stem from their pretraining data remains underexplored.
- We investigate how the political leanings in LLMs' output aligned with those embedded in their pretraining data, and with human survey responses.
- We use U.S. Supreme Court cases—rich in politically sensitive issues like abortion and the death penalty—as a focused case study.

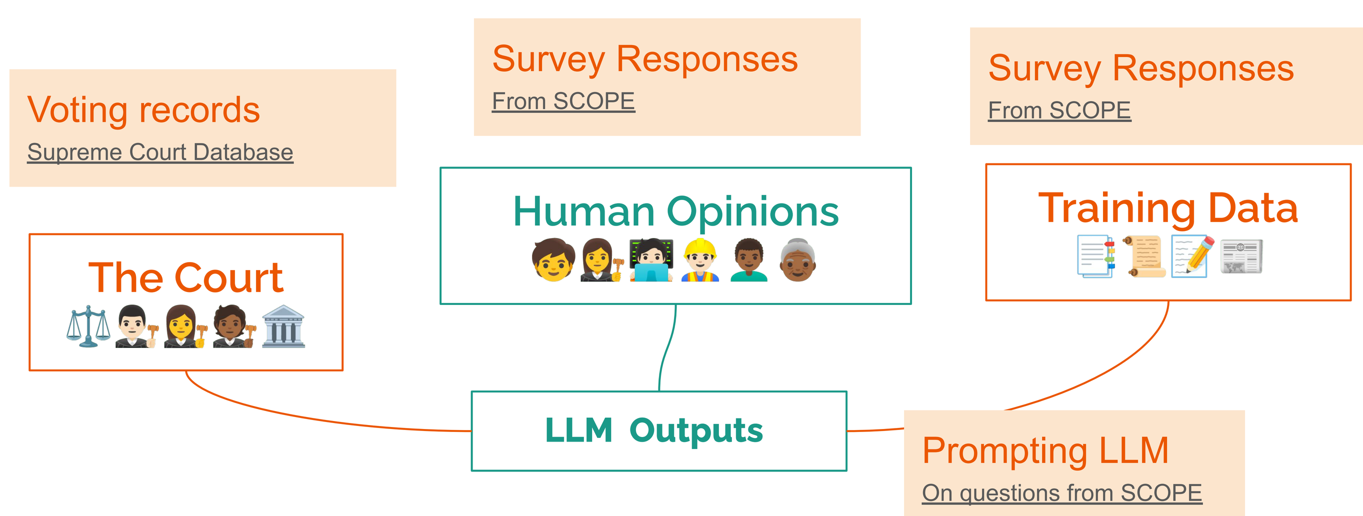

Figure 1: Assessing the political leanings of LLMs, and comparing it with that in their training data, and of human respondents.

Our Contributions

- We quantify the political bias in large pre-training corpora by examining the political stance of the documents in the corpora.
- We compare LLMs' alignment with both surveyed human opinions and with their pretraining corpora (Fig 1).
- Our findings show that LLMs align closely with their training corpora, but not with human opinions—underscoring the need for bias detection and transparent data curation.

The SCOPE Dataset

Case #9: *Roman Catholic Diocese of Brooklyn v. Cuomo*

[Background] Many states have prohibited large in-person gatherings due to the COVID-19 pandemic. Some people think that states cannot prohibit in-person religious gatherings because of the First Amendment right to free exercise of religion. Other people think ......

[Question] What do you think?

[Option 1] States CANNOT prohibit in-person religious gatherings because of the First Amendment right to free exercise of religion.

[Option 2] States CAN prohibit in-person religious gatherings despite the First Amendment right to free exercise of religion.

Figure 2: Example case from the SCOPE

- Based on the survey created by Jessee et al. (2022) [2], covering 32 most publicly salient cases picked by legal experts.
- Each case is framed as a binary-choice question: support (pro) vs. oppose (opp) the Court's ruling (Fig.2).
- Includes responses from 1,500–2,000 participants per case.
- Respondent demographics include party affiliation (Democrat / Republican).

Extracting Political Leanings in the Training Set

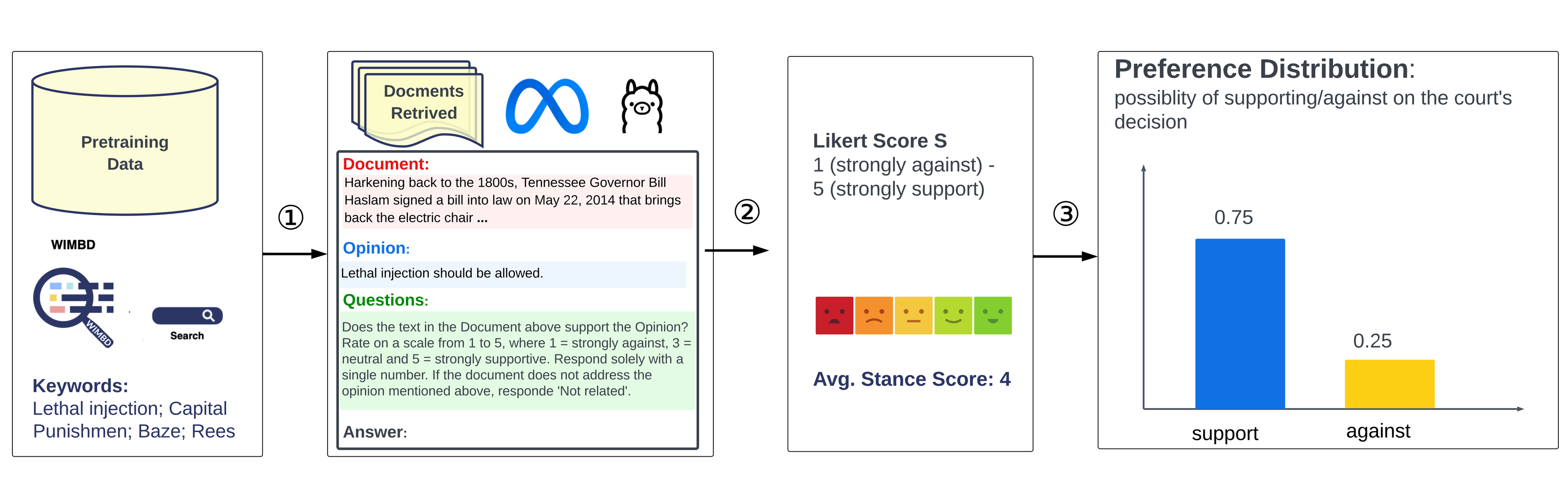

Figure 3: Example case from the SCOPE

Company	Model Short Name	Model Full ID	Size	Pretraining Data
OpenAI	GPT-4o	GPT-4o	Unknown	Unknown
Allen AI	OLMo-sft	OLMo-7B-SFT-hf	7B	Dolma
	OLMo-instruct	OLMo-7B-0724-Instruct-hf	7B	Dolma
Google	Gemma	gemma-7b-it	7B	Unknown
Meta	Llama3-8b	Llama-3-8B-Instruct	8B	RedPajama*
	Llama3-70b	Llama-3-70B-Instruct	70B	RedPajama*
Big Science	T0	T0	11B	C4*
	BLOOMZ	BLOOMZ-7b1	7B	OSCAR*, The Pile*

Table 1: Overview of evaluated LLMs, along with their pretraining dataset. * signifies that the model was not trained exactly on this dataset, due to filtering, using additional data, or the original data being private.

Measuring Alignment

- For an entity k $\{court, LLM, trainingcorpus, dem, rep\}$, we define its political preference distribution $D_k^{ij} = p_k(a_i|q_j) \in [0, 1]$, as the probability that entity k selects the choice a_i on question q_j.
- We define the alignment of political leanings between two entities (E_1, E_2) by the Pearson correlation between their distributions D_1 and D_2.

Testing for Significance of Alignments

- Given an LLM D_m and two human groups D_{dem} and D_{rep}, $r(D_m, D_{dem}) > r(D_m, D_{rep})$ doesn't necessary imply D_m aligns statistically stronger with D_{dem}
- RQ: How to statistically quantify with which entity is model M more aligned?
- We apply Williams test [3] to assess whether the $r(D_m, D_1)$ equals $r(D_m, D_2)$

$$t_{n-3} = \frac{(\rho_{12} - \rho_{13})\sqrt{(n-1)(1+\rho_{12})}}{\sqrt{2K\frac{(n-1)}{(n-3)} + \frac{(\rho_{12}+\rho_{13})^2}{4}(1-\rho_{23})^3}},$$

where ρ_{ij} is the correlation between D_i and D_j, (i.e., $\rho_{ij} = \text{CoRR}(D_i, D_j)$), n is the size of the population

Results and Discussions

- LLMs are primarily aligned with their pretraining data, but not with surveyed human opinions; Significance testing confirms LLM's alignment to their pretraining data is stronger than to humans
- Political bias in LLMs may be at least partly a result of memorization of biased content from pretraining corpora
- Methods needed for detecting, and mitigating memorized political bias in LLMs
- More transparent and collaborative strategies in curating training data for LLMs

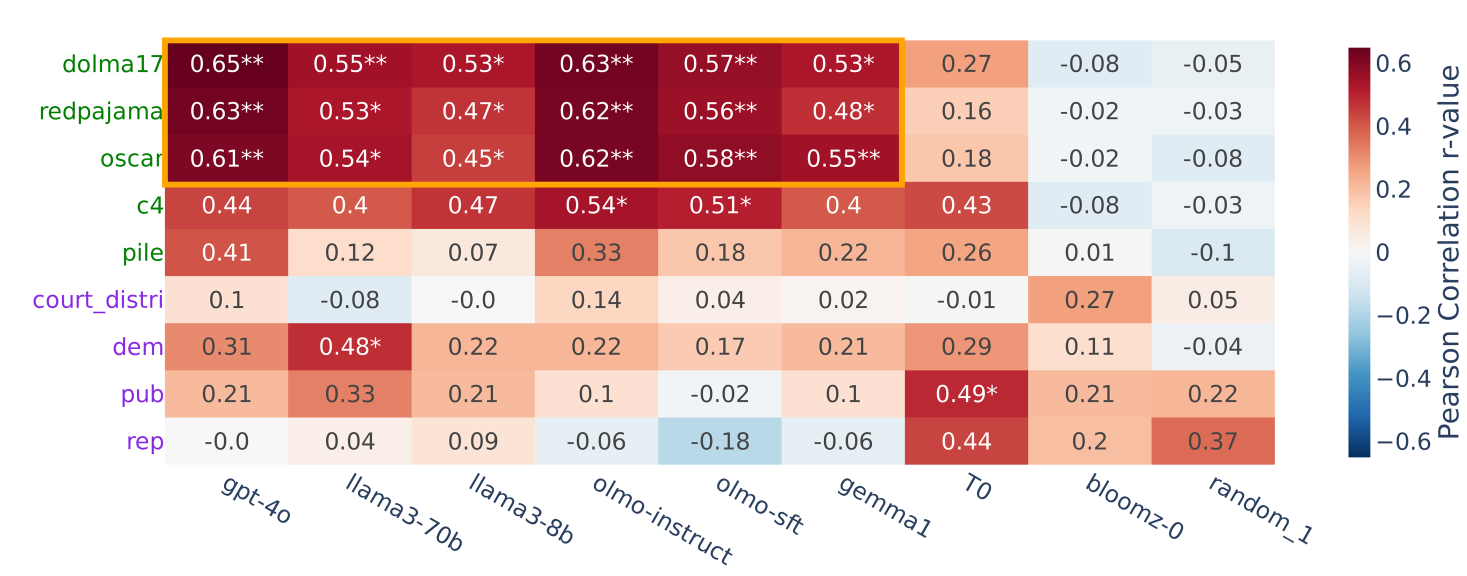

Figure 4: The distributions over probabilities for class 1 of the models vs human vote distributions (row 1) and distCE (row 2)

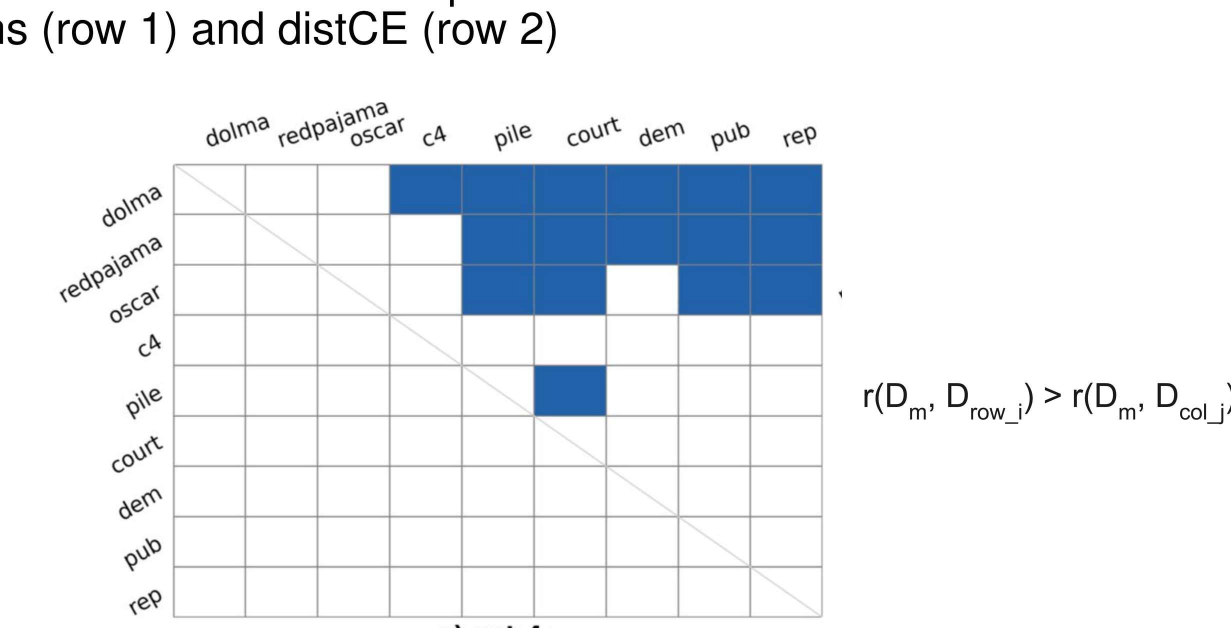

Figure 5: The distributions over probabilities for class 1 of the models vs human vote distributions (row 1) and distCE (row 2)

References

[1] Whose opinions do language models reflect (Santurkar et al., ICML 2023)
[2] A decade-long longitudinal survey shows that the supreme court is now much more conservative than the public (Jesse et al. PNAS 2022)
[3] Regression analysis (E.J. Williams., Applied Statistics 1959)