Improving the Distributional Alignment of LLMs using Supervision

Gauri Kambhatla, Sanjana Gautam, Angela Zhang, Alex Liu,
Ravi Srinivasan, Junyi Jessy Li, Matthew Lease

The University of Texas at Austin

GOOD SYSTEMS
Ethical AI at UT Austin
goodsystems.utexas.edu

In this work, we show that simple supervised calibration can consistently improve the alignment of LLM-generated distributions with diverse human population groups, as measured across three datasets using multiple elicitation methods.

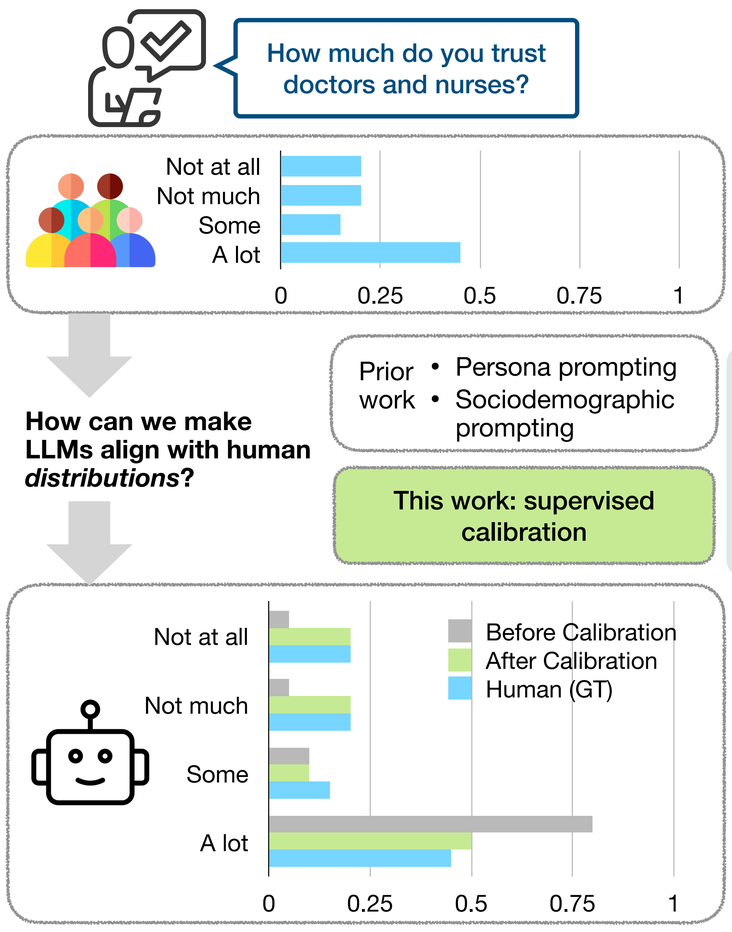


Figure 1: Overview. We elicit distributions from LLMs and calibrate them to better align with human response distributions.

Why is this important?

Prior work studying opinion alignment using common prompting techniques (1) yields **inconsistent results** across tasks, models, and elicitation methods and/or (2) **doesn't model in-group disagreement**

Benchmark

Survey datasets	LLM family: models	Probability elicitation	Prompt
Wellcome Global Monitor (WGM)	Claude: 3, 3.5v1, 3.5v2	Verbalized	With SD variables
OpinionQA (OQA)	Mistral: small, large	Self-random [Xiong et al., 2024]	Without SD variables (Base)
World Values Survey (WVS)	Llama: 3-70B, 3.1-70B, 3.2-1B, 3.2-11B, 3.2-90B	Paraphrase [Xiong et al., 2024]	
	Qwen: 2.5-72B		
	OLMo-2: 7B, 7B-SFT, 7B-DPO, 7B-Instruct		

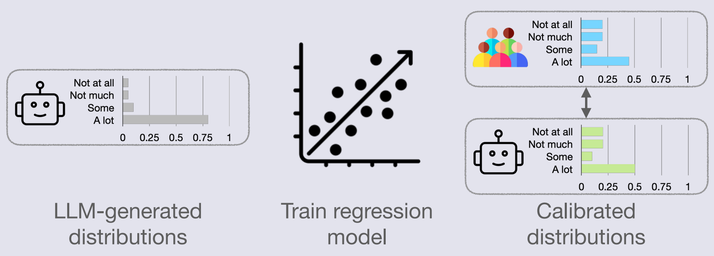
Preliminaries

- **Sociodemographic (SD) prompting:** defining a persona with demographic variables within a prompt, with the goal of simulating their opinions or behaviors.
- **Distributions**: Parameterized as distributions over answer choices to the ordinal survey questions, e.g., [0.2, 0.2, 0.15, 0.45] as shown in Figure 1
- **Evaluation**: We use the Opinion Alignment metric [Santurkar et al., 2023] to measure the similarity between an elicited distribution and ground truth distribution

Supervised Calibration & Alignment

Calibration

We apply supervised **regression** to transform LLM-generated distributions to align them with the human distributions



LLM-generated distributions | Train regression model | Calibrated distributions

- Settings: dataset-LLM-elicitation method set
- We learn a separate regression model for each setting

Calibration effect: less variance in alignment between LLMs and elicitation methods

Figure 2: Std dev vs. opinion alignment. Each point represents the average alignment for each setting. **Calibration decreases variance in opinion alignment and standard deviation *between* settings.**

	Model	Base prompt P	Base prompt P_C	Base prompt S	Base prompt S_C	Base prompt V	Base prompt V_C	Sociodemographic prompt P	Sociodemographic prompt P_C	Sociodemographic prompt S	Sociodemographic prompt S_C	Sociodemographic prompt V	Sociodemographic prompt V_C
WGM	Claude-3.5-v2	66.2	**85.2**	59.5	**85.1**	89.3	87.7	65.4	**84.4**	61.4	**84.2**	89.0	89.3
	Llama-3.2-90B	68.1	**84.6**	73.0	**86.2**	84.8	**89.0**	70.6	**86.0**	67.6	**86.1**	85.0	**89.8**
	Mistral-large	62.4	**84.7**	72.0	**83.9**	89.4	88.9	68.4	**84.7**	63.0	**84.9**	87.3	88.4
	OLMo-2-7B-I	59.6	**81.5**	67.7	**80.4**	59.8	**85.1**	62.3	**82.5**	64.5	**82.7**	69.8	**84.6**
	Qwen-2.5-72B	57.7	**83.1**	53.3	**83.2**	88.2	87.1	66.0	**84.1**	63.4	**85.2**	89.1	89.4
	Average	62.8	**83.8**	65.1	**83.8**	82.3	**87.6**	66.5	**84.3**	64.0	**84.6**	84.0	**88.3**
OQA	Claude-3.5-v2	70.5	**88.8**	72.5	**90.4**	91.7	91.7	76.1	**89.9**	73.3	**89.6**	91.9	91.6
	Llama-3.2-90B	79.3	**89.6**	75.3	**89.4**	86.8	87.9	79.2	**90.1**	76.1	**90.0**	83.4	**85.9**
	Mistral-large	79.5	**89.9**	75.3	**88.3**	85.0	86.2	75.8	**89.2**	72.4	**89.5**	83.8	84.7
	OLMo-2-7B-I	72.6	**89.0**	72.3	**88.4**	65.4	**79.9**	72.8	**88.4**	70.5	**88.6**	68.5	**81.5**
	Qwen-2.5-72B	73.9	**89.7**	67.0	**88.8**	88.4	87.9	74.4	**90.0**	71.3	**89.5**	89.2	88.6
	Average	75.2	**89.4**	72.5	**89.1**	83.5	**86.7**	75.7	**89.5**	72.7	**89.4**	83.4	**86.5**
WVS	Claude-3.5-v2	46.5	**80.3**	51.0	**80.3**	75.2	**80.3**	61.0	**80.4**	56.8	**80.4**	75.6	81.7
	Llama-3.2-90B	61.6	**79.9**	59.1	**80.3**	64.6	**80.3**	62.1	**80.8**	59.5	**81.5**	67.7	**82.7**
	Mistral-large	48.8	**80.3**	44.0	**82.2**	72.8	**80.3**	54.3	**80.4**	51.5	**80.3**	76.6	**83.8**
	OLMo-2-7B-I	75.3	**80.3**	74.9	**80.3**	86.6	86.5	58.3	**79.2**	60.0	**77.2**	86.0	**89.8**
	Qwen-2.5-72B	39.6	**82.0**	42.4	**82.3**	74.0	77.9	49.1	**82.4**	49.4	**81.5**	73.4	**81.7**
	Average	54.4	**80.6**	54.3	**81.1**	74.6	**81.1**	57.0	**80.6**	55.4	**80.2**	75.9	**83.9**
Average	Claude-3.5-v2	61.1	**84.8**	61.0	**85.3**	85.4	**86.6**	67.5	**84.9**	63.8	**84.7**	85.5	**87.5**
	Llama-3.2-90B	69.7	**84.7**	69.1	**85.3**	78.7	**85.7**	70.6	**85.6**	67.7	**85.9**	78.7	**86.1**
	Mistral-large	63.6	**85.0**	63.8	**84.8**	82.4	**85.1**	66.2	**84.8**	62.3	**84.9**	82.6	**85.6**
	OLMo-2-7B-I	69.2	**83.6**	71.6	**83.0**	70.6	**83.8**	64.5	**83.4**	65.0	**82.8**	74.8	**85.3**
	Qwen-2.5-72B	57.1	**84.9**	54.2	**84.8**	83.5	**84.3**	63.2	**85.5**	61.4	**85.4**	83.9	**86.6**
	Average	64.1	**84.6**	64.0	**84.6**	80.1	**85.1**	66.4	**84.8**	64.0	**84.7**	81.1	**86.2**

Table 1: Opinion alignment before and after calibration for each dataset, LLM, and elicitation method. Each pair of columns compares the base-generated or SD generated distributions to the calibrated distributions (C) for each elicitation method: paraphrase (P), self-random (S), and verbalized (V). **Bolded** values are significant between each pair. The best performance for each model is highlighted in yellow. The best averages across models and datasets are highlighted in blue.

Findings: Alignment

(RQ1) Is SD prompting effective for distributional alignment across datasets?

- No single LLM or elicitation method delivers reliably consistent results across datasets
- Affirms prior work in our controlled & systematic benchmark setting with distributions

(RQ2) Can post-hoc calibration improve alignment?

- Calibrated alignment scores are **higher in 275/290 (94.8%)** settings (across all datasets, models, and elicitation methods) by an average of **16.3%**
- Calibrated alignment scores have lower standard deviation in **253/290 (87.2%)** settings which is on average **1.62 times** lower.

Calibration effect: *consistent* improvement in opinion alignment across LLMs, datasets, and methods

Findings: Individual Groups

World Region (WGM)	Base & Verb.	+ Calibration
Aus/NZ	85.6	82.6
Central Africa	69.7	**81.1**
Cent. America & Mexico	74.1	**84.9**
Central Asia	84.7	82.5
Eastern Asia	82.4	85.0
Eastern Africa	81.1	87.5
Middle East	84.3	89.3
North Africa	79.3	85.6
Northern America	84.8	84.8
Northern Europe	86.2	84.1
South America	73.3	**83.9**
South Asia	83.5	84.6
Southern Africa	74.4	**85.9**
Southern Europe	81.9	87.5
Western Africa	80.6	**89.4**
Western Europe	86.3	85.3

(RQ3) How does calibration affect individual groups?

Calibration improves alignment most for sociodemographic groups for which LLMs are aligned with least. For world region, these tend to be Global South regions. See other demographics in our paper.

Table 2: Opinion alignment scores for base-prompted verbalized elicitation **with and without calibration** on the world region demographic of the WGM dataset. Bolded values indicate significant difference.

Findings: Minimal supervision

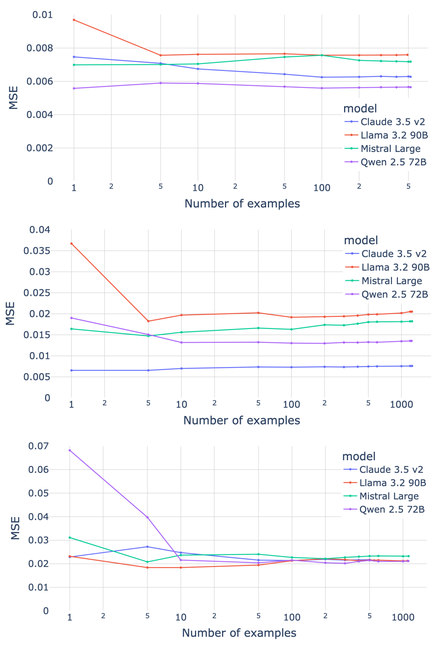


Figure 3: MSE vs. training example plots for WGM, OQA, and WVS

(RQ4) How many supervised training examples do you need?

As few as 1-10 gold examples can be used to calibrate distributions. Figure 3 shows Mean Squared Error (MSE) of the regression models for four LLMs on each dataset. The regression models converge with 1-10 training examples depending on model and dataset.

Paper

Code/Data

@gkambhat
gkambhat@utexas.edu
gaurikambhatla.github.io