Political Discussion and Information Transmission Under Social Pressure

Xavier Bauman, Brody Mills; Marli Dunietz

Overview

The polarization of political discussion in the United States has resulted in an information environment in which someone's sincerity often depends on their perception of how others will react to their opinion.

As a result, American audiences following public discourse may be conscious of how people distort their opinions in a socially acceptable direction and try to infer what their "real" opinion is. Errors of inference can occur on either side of information transmission. We ask how people express their opinions and how accurately others infer the beliefs behind that expression.

We examine how social norms and incentives influence opinion expression and interpretation. Our focus will rest upon the cost and benefits of encouraging personal versus impersonal language, and how perceptions of extremity and polarization are affected. We collected responses to various policy proposals from a random population sample. Another group then graded responses based on what they estimated the original respondents believed.

We use LLMs to build a rubric for sentiment, emotional intensity, and argument quality, enabling consistent large-scale annotation of the responses. We hypothesize that certain response characteristics such as argumentative strength and emotional intensity shape how accurately others interpret original respondents' beliefs.

Methodology

- **First**, we annotated 2 large datasets from a survey in which encoders responded to policy proposals. Other participants then decoded those responses to find the original encoders' beliefs. We annotated the encoders responses by grading them as either Neutral (N), Moderate (M), or Strong (S).
- **Next**, we used Gemini 3 Pro to annotate these same datasets, defining a codebook that would allow the LLMs to consistently replicate our annotations on a large scale across 6 issue areas: the death penalty, gender transition surgery, homelessness, immigration, renewable energy, and abortion.
- **Finally**, we plan to use these annotations to determine how the characteristic of response strength determines decoder accuracy.

LLM Annotation Training

PROMPT 1:
LLM Training Set. Emulate Human Annotation (Ann.).

PROMPT 2:
Compare and Categorize Response Types.

PROMPT 3:
LLM – Human Ann. Comparison and Explanation of Reasoning.

PROMPT 4:
LLM Predicts Human Ann for Next Response Set.

PROMPT 5:
LLM – Human Ann. Comparison and Explanation of Reasoning.

Prompt 6:
Construct Rubric For Predicting Human Ann.

Prompt 7:
LLM Predicts Human Ann. For Last Set.

Results

- As described in **Figure 1**, Gemini 3 Pro was able to consistently replicate human annotation overall across most issue areas.
- The LLM was **unusually INCONSISTENT** with replicating human annotation within the **immigration** issue area.
- The LLM was **unusually CONSISTENT** with replicating human annotation within the **renewable energy** issue area.

Figure 1

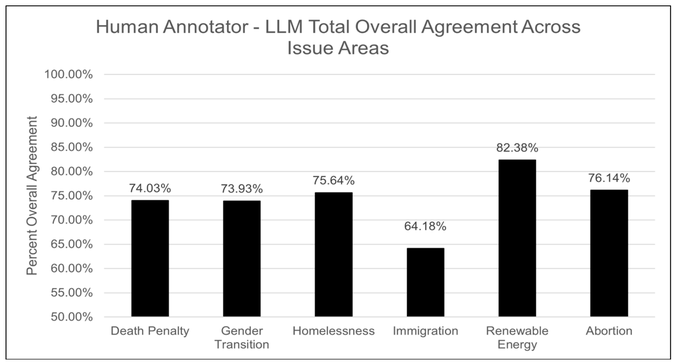
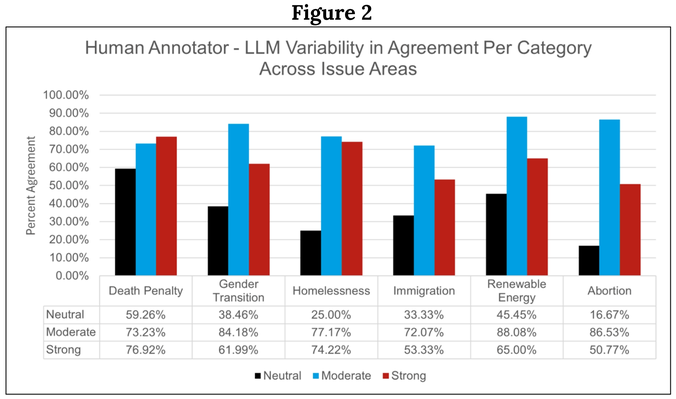


Figure 2



	Death Penalty	Gender Transition	Homelessness	Immigration	Renewable Energy	Abortion
Neutral	59.26%	38.46%	25.00%	33.33%	45.45%	16.67%
Moderate	73.23%	84.18%	77.17%	72.07%	88.08%	86.53%
Strong	76.92%	61.99%	74.22%	53.33%	65.00%	50.77%

- **Figure 2:** Across most issue areas, the LLM best replicates human performance for **moderate** responses.
- Within **renewable energy**, it is unusually **consistent**.
  - This is likely due to this issue having more utilitarian responses and relatively less emotional attachment than other issues
- The LLM is the least consistent with **neutral** responses.
  - It often struggled to exclude responses with unrelated information

Figure 3

Death Penalty

	LLM: N	LLM: M	LLM: S
Ann1: N	16	8	3
Ann1: M	11	186	57
Ann1: S	4	44	160

Immigration

	LLM: N	LLM: M	LLM: S
Ann1: N	11	19	3
Ann1: M	27	271	78
Ann1: S	4	80	96

Gender Transition

	LLM: N	LLM: M	LLM: S
Ann1: N	15	20	4
Ann1: M	23	330	39
Ann1: S	3	81	137

Renewable Energy

	LLM: N	LLM: M	LLM: S
Ann1: N	5	5	1
Ann1: M	4	170	19
Ann1: S	0	14	26

Homelessness

	LLM: N	LLM: M	LLM: S
Ann1: N	2	5	1
Ann1: M	35	365	73
Ann1: S	1	57	167

Abortion

	LLM: N	LLM: M	LLM: S
Ann1: N	1	5	0
Ann1: M	14	167	12
Ann1: S	0	32	33

- If the LLM is **consistent** with human performance, most of the data should be concentrated within the bolded diagonal cells.
  - Most data is concentrated within the **moderate** agreement cell (center), showing it is most consistent with **moderate** responses.
  - There is less data in the **strong** agreement cell, demonstrating weaker consistency with **strong** responses.

Discussion

For most issue areas, Gemini 3 Pro was able to replicate human annotation consistently. It was **unusually consistent** with the **renewable energy** issue area, and **unusually inconsistent** within the **immigration** issue area. We believe this is influenced by the divisiveness and emotional attachment of these issue areas.

- Immigration is a highly salient, divisive, and emotionally charged topic. It is also one in which background context contributes significantly to whether a response is moderate or strong.
  - The LLM frequently misidentified statements with nativist undertones as moderate, likely a result of human annotators having access to background information which allows for inferences the LLM doesn't make often.
- Renewable energy is relatively less salient and divisive than the other issue areas, demonstrated by the larger portion of detached, moderate responses.
  - Because of the prevalence of these responses, we believe the higher consistency is the result of the LLM being better at categorizing logical arguments without underlying emotional sentiments or moral components.

Issues like the **death penalty** and **homelessness** are emotionally charged, but often the strong responses have emotional sentiments and moral components that are easier to differentiate from moderate responses.

- Strong responses often contain absolute moral arguments within these issue areas, and moderate arguments are usually of a utilitarian or non-interventionist nature.

We believe that **lack of background context**, **issue saliency**, and **emotional sentiment** are major contributors to the variability in consistency across results and the disparity between LLM and human performances.

Overall, the LLM was most consistent at replicating human performance across moderate responses, and the least consistent with replicating human performance across neutral responses. We expect this to be the result of the qualities about political discourse within specific issue areas.

Next Steps

- **First**, we need to reach a high level of LLM agreement with human annotations. While Gemini 3 Pro was able to produce accurate annotations for some issues, we will continue to train the LLM in order to produce more accurate, context-aware annotations that effectively isolate specific aspects of responses.
- **Next**, we plan to use the higher-trained LLM to consistently annotate responses on a large scale.
- **Finally**, we search for a relationship between certain aspects of speech in responses and the accuracy of human decoders interpreting those responses. We predict that we will find that certain aspects of speech will correlate with information distortion.

Implications

Our findings will contribute to the literature on how citizens learn about public opinion, and how distortions can arise within public discourse. By determining how aspects of speech like communication style and emotional intensity affect individual perceptions, we can illuminate where these distortions may be originating and improve the accuracy of information transmission.

References

- Bojić, L., Zagovora, O., Zelenkauskaite, A., Vuković, V., Čabarkapa, M., Jerković, S. V., & Jovančević, A. (2025). Comparing large Language models and human annotators in latent content analysis of sentiment, political leaning, emotional intensity and sarcasm. *Scientific Reports*, *15*(1), 11477. https://doi.org/10.1038/s41598-025-96508-3
- Braghieri, L. (2024). Political Correctness, Social Image, and Information Transmission. *American Economic Review*, *114*(12), 3877–3904. https://doi.org/10.1257/aer.20210039
- Huang, Y. (n.d.). Breaking the Spiral of Silence [Harvard University]. In *Yihong Huang - Ph.D. Candidate in Economics*. https://sites.harvard.edu/yihong-huang/job-market-paper/
- Modarressi, I., Spiess, J., & Venugopal, A. (n.d.). Causal Inference on Outcomes Learned from Text. *Working Paper*. https://arxiv.org/pdf/2503.00725
- Törnberg, P. & Institute for Language, Logic and Computation (ILLC), University of Amsterdam. (2024). Best Practices for Text Annotation with Large Language Models. *Institute for Language, Logic and Computation (ILLC), University of Amsterdam*. https://arxiv.org/pdf/2402.05129