Appendix A

Previous research[1] demonstrated the potential of employing generalized large language models (LLMs) for analyzing text data. This study explores whether LLMs can be used to correctly identify the use of bipartisan signals. We evaluated the performance of ChatGPT 3.5 Turbo, ChatGPT 4o, Claude 3 Opus, Claude 3.5 Sonnet. We found the accuracy of these LLMs to be sufficiently low, so we resorted to using human coders. Below we expand on the prompts used and the corresponding results.

Method

In order to detect bipartisan signals, one needs to identify the political party of the politician and whether the contents of the tweets were congruent with the party's stance on the issue. The twitter dataset contained political parties of politicians. Hence, the LLMs task was to determine congruency of tweet with the party's position.

Specifically, we wanted the LLMs to categorize the tweets into one of three categories: 0 for congruency with the party's values, 1 for incongruency (i.e. bipartisan signals), and **2** for irrelevance (e.g. holiday celebrations).

We used various prompts, and the results remained sufficiently inaccurate regardless of the prompt. Below we focus on the results obtained from the prompts that yielded the highest accuracy. The prompt read:

```
Given a tweet and the author's party affiliation (Democrat or
Republican), assign one of these labels:
0: The message reflects typical positions or rhetoric associated with
the author's party affiliation 1: The message contradicts or
challenges typical positions associated with the author's party
affiliation 2: The message is politically neutral or focused on
general updates, celebrations, or acknowledgments
Consider both the explicit content and underlying tone/implications
when assessing party alignment. Focus on whether the message would be
generally expected from someone of that party affiliation in
contemporary American politics.

please return in the form of a csv without altering the tweets itself.
please label all the tweets
```

---

[1] Rathje, S., Dan-Mircea Mirea, Ilia Sucholutsky, Raja Marjieh, Robertson, C. E., & Van, J. J. (2024). GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences*, *121*(34). https://doi.org/10.1073/pnas.2308950121

Method

The table summarizes the confusion matrix by LLM used. As seen below, all LLMs struggled with classifying bipartisan signals. For example, ChatGPT 4o only accurately classified 55.6% that were deemed bipartisan signals by human coders. While ChatGPT 3.5 Turbo was able to classify 100% of the bipartisan signals, it also greatly misidentified congruent and irrelevant tweets as bipartisan. As another example, Claude 3.5 Sonnet only accurately classified 33% that were deemed bipartisan signals by human coders.

Metric	Macro Precision	Macro Recall	Macro F1	Accuracy
Claude 3 Opus	0.515	0.494	0.445	0.614
Claude 3.5 Sonnet	0.593	0.622	0.581	0.729
ChatGPT 3.5 Turbo	0.490	0.378	0.089	0.083
ChatGPT 4o	0.518	0.606	0.460	0.573

Claude 3 Opus

Model Prediction	Ground Truth		
congruent	0.87	0.67	0.50
bipartisan	0.06	0.22	0.05
irrelevant	0.07	0.11	0.45

Claude 3.5 Sonnet

Model Prediction			
congruent	0.93	0.56	0.38
bipartisan	0.01	0.33	0.02
irrelevant	0.06	0.11	0.60

ChatGPT 3.5 Turbo

Model Prediction				
	congruent	0.04	0.00	0.03
	bipartisan	0.96	1.00	0.88
	irrelevant	0.00	0.00	0.10

ChatGPT 4o

Model Prediction				
	congruent	0.92	0.33	0.62
	bipartisan	0.03	0.56	0.04
	irrelevant	0.05	0.11	0.34