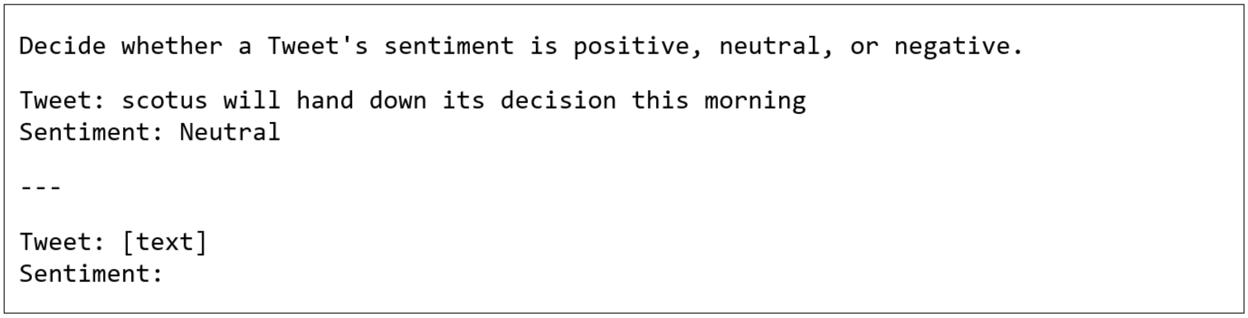
A Appendix A: Additional Tables and Figures

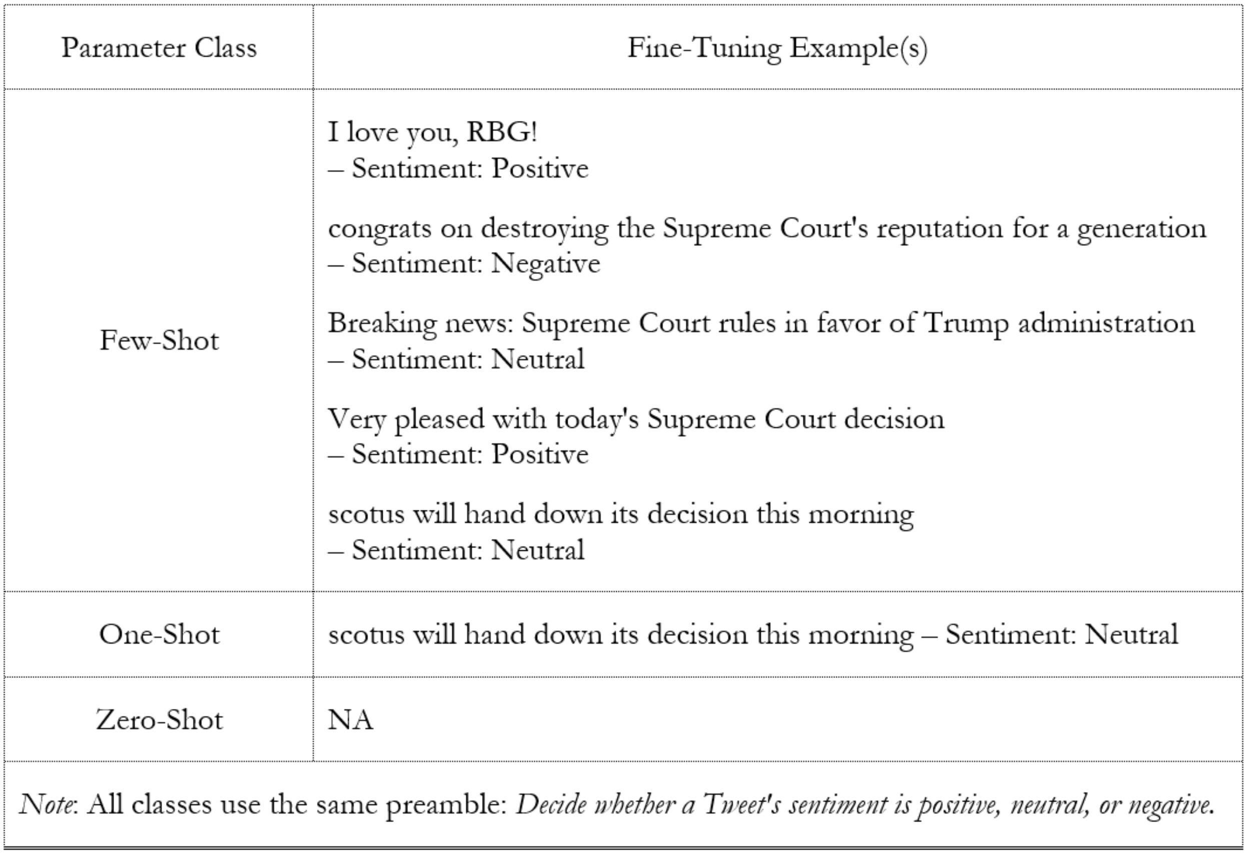
To explore how the size and complexity of the model affects its classification performance, we iteratively tested multiple GPT-3 variants (2.7 billion, 7 billion, 13 billion, and 175 billion parameters) using prompts structured as follows:



```
Decide whether a Tweet's sentiment is positive, neutral, or negative.

Tweet: scotus will hand down its decision this morning
Sentiment: Neutral

---

Tweet: [text]
Sentiment:
```

In addition to varying the number of model parameters, we vary whether the prompt is zero-shot, one-shot, or few-shot, using the examples listed in Table A1.

Table A1. Few-Shot Examples for Sentiment Classification Task



Parameter Class	Fine-Tuning Example(s)
Few-Shot	I love you, RBG! – Sentiment: Positive congrats on destroying the Supreme Court's reputation for a generation – Sentiment: Negative Breaking news: Supreme Court rules in favor of Trump administration – Sentiment: Neutral Very pleased with today's Supreme Court decision – Sentiment: Positive scotus will hand down its decision this morning – Sentiment: Neutral
One-Shot	scotus will hand down its decision this morning – Sentiment: Neutral
Zero-Shot	NA

Note: All classes use the same preamble: *Decide whether a Tweet's sentiment is positive, neutral, or negative.*

Figure A1 reports performance across every combination of model and prompt variant. As expected, larger models generally outperform smaller models, and providing more examples in the prompt consistently improves performance. The smallest model variants (2.7b and 6.7b parameters) perform quite poorly, requiring few-shot learning before their output is even modestly correlated with the expert scores. But the two largest variants perform well regardless of prompt design choices. The 175-billion parameter GPT-3 predicts whether a tweet was negative or positive in 87.5% of cases with one-shot learning, and 88.4% of the time with few-shot learning. The 13-billion parameter variant performs nearly as well, with accuracies of 85.1% for one-shot learning and 87.4% for few-shot learning. Choosing a less-capable variant of the model may be advantageous for some researchers, since (as of writing) these models are only available through OpenAI's paid Application Programming Interface (API), and the per-token rates for smaller models are less expensive. For a dataset this size, however, the costs were minimal. In October 2022, coding our 945 tweets cost $4.86 with one-shot prompting and 175b parameters, versus $0.46 with 13b parameters. By October 2024, the per-token costs for similar models had been reduced nearly twenty-fold.

Figure A1. GPT-3 performance by at sentiment classification task, by prompt and model variant (Ada = 2.7B, Babbage=6.7B, Curie=13B, Davinci = 175B)

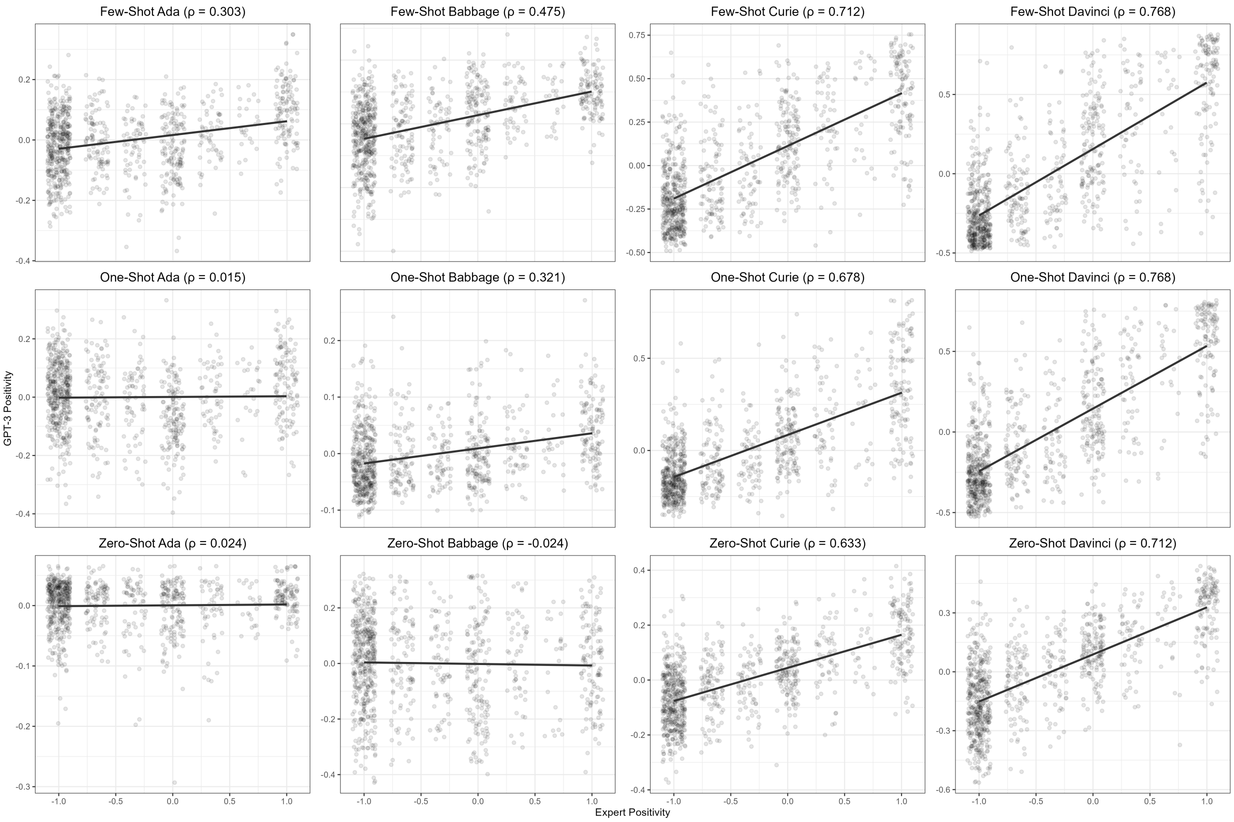


Table A2. Prompts for Twitter sentiment application

For the tweets following the *Masterpiece Cakeshop* decision:

Read these tweets posted the day after the US Supreme Court ruled in favor of a baker who refused to bake a wedding cake for a same-sex couple. For each tweet, decide whether its sentiment is Positive, Neutral, or Negative.

Tweet: #SCOTUS set a dangerous precedent today. Although the Court limited the scope to which a business owner could deny services to patrons, the legal argument has been legitimized that one's subjective religious convictions trump (no pun intended) #humanrights. #LGBTQRights

Sentiment: Negative

Tweet: Thank you Supreme Court I take pride in your decision!!!! #SCOTUS

Sentiment: Positive

Tweet: Supreme Court rules in favor of baker who would not make wedding cake for gay couple

Sentiment: Neutral

Tweet: Supreme Court rules in favor of Colorado baker! This day is getting better by the minute!

Sentiment: Positive

Tweet: Can't escape the awful irony of someone allowed to use religion to discriminate against people in love. Not my Jesus. #opentoall #SCOTUS #Hypocrisy #MasterpieceCakeshop

Sentiment: Negative

Tweet: I can't believe this cake case went all the way to #SCOTUS . Can someone let me know what cake was ultimately served at the wedding? Are they married and living happily ever after?

Sentiment: Neutral

Tweet: [text]

Sentiment:

For tweets following the *Mazars* decision:

Read these tweets posted the day after the US Supreme Court ruled that sitting presidents are not immune to state criminal subpoenas, and that President Trump was obliged to disclose his tax returns to the Manhattan District Attorney. For each tweet, decide whether its sentiment is Positive, Neutral, or Negative.

Tweet: SCOTUS just ruled Manhattan DA CAN get trumps financials and tax returns. This is a great day for the ruke of law and America.

Sentiment: Positive

Tweet: Justice #ClarenceThomas is waste of space on the #scotus

Sentiment: Negative

Tweet: BREAKING: Supreme Court Justice Ruth Bader Ginsburg has been hospitalized for a possible infection, per a SCOTUS spokesperson. @Scotus @ruthbadergins

Sentiment: Neutral

Tweet: Today the Supreme Court let @realDonaldTrump know that he is not above the law!

Sentiment: Positive

Tweet: The Supreme Court is going to disappoint us tomorrow. And trump will feel even more untouchable. He'll brag about it at his Klan rallies. Sweaty orange spray tan pooling above his lip, smug faced as he gloats and brags. It makes me sick.

Sentiment: Negative

Tweet: Both SCOTUS rulings in Trump financial records sent back to lower courts. Practically speaking that means no turnover of records immediately in either case. #7News

Sentiment: Neutral

Tweet: [text]

Sentiment:

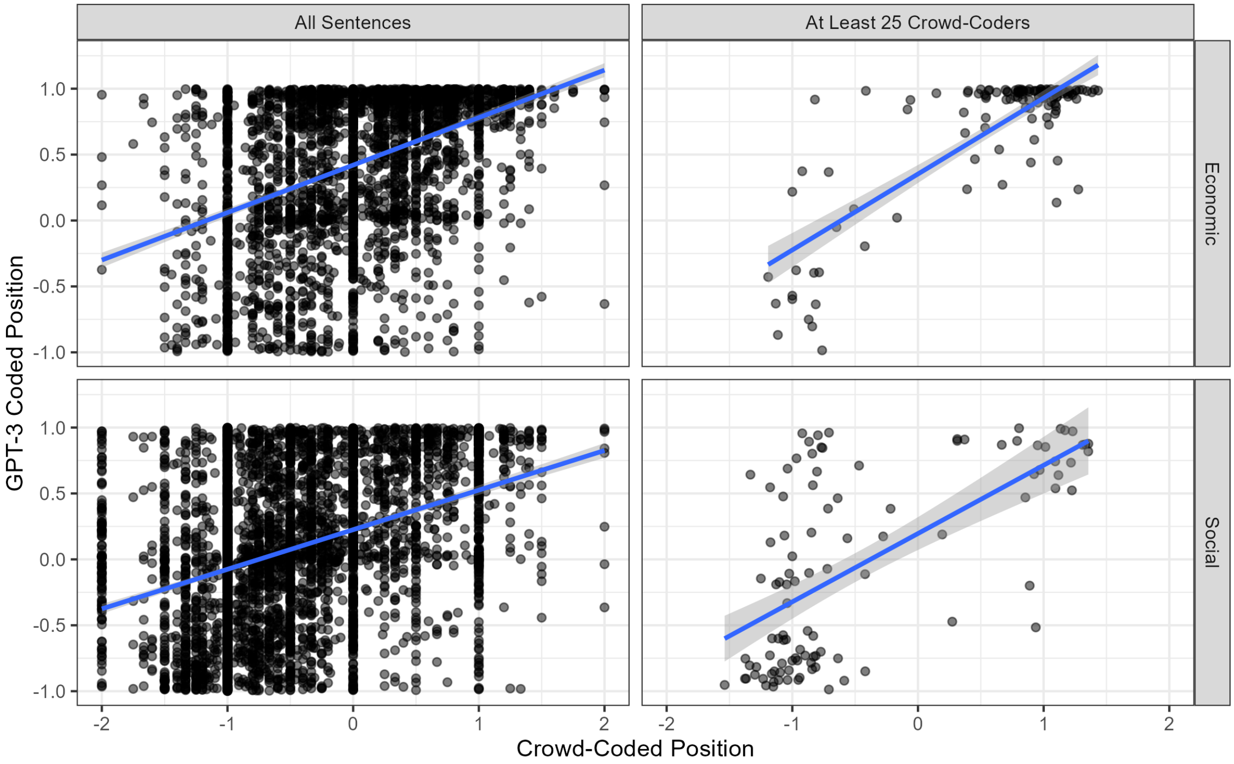
Table A3. Sample of political ads where GPT-3 and expert classifications disagreed

Text	Expert Label	GPT-3 Label
[Announcer]: 12 years ago Susan Collins made a pledge. [Collins]: "I have pledged that if I'm elected I will only serve two terms, regardless of whether terms limits law - constitutional amendment passes or not. Twelve years is long enough to be in public service, make a contribution and then come home and let someone else take your place. Twelve years is long enough to be in public service, make a contribution and then come home and let someone else take your place." [Allen]: "I'm Tom Allen and I approved this message." [PFB]: TOM ALLEN FOR SENATE	Negative	Positive
[Norm Coleman]: "Here's somethings you're probably going to see some more of from the other side. First, they'll show you a crummy picture, bad hair day. Then, they'll play some scary music. They'll say I'm in the pocket of lobbyists, special interests, but I fought for ethics reform to restore trust in Congress. They'll say I'm a rubber stamp for George Bush even though the Washington Post has ranked me as one of the most independent Senators. I'm Norm Coleman, I approve this message, because I just thought you should be prepared. Ouch, where'd they get that?" [PFB]: COLEMAN FOR SENATE '08	Positive	Negative
[Andrew Rice]: "I'm Andrew Rice. My faith teaches to help those in need. That's why I served as a Christian missionary. After my brother David was killed in the World Trade Center on 9/11, I ran for public office to change things. Now I'm running for U.S. Senate because Washington isn't solving our problems. Jim Inhofe's been in Washington 22 years and he's lost his way. I'm Andrew Rice I approve this message because it's time for leadership we can have faith in...again." [PFB]: ANDREW RICE FOR U.S. SENATE	Negative	Positive

Table A4. Sentence-level correlation between one-shot GPT-3 ideology score and crowd-coders in manifesto application

Policy	Number of Crowd-Coders	Correlation	$N_{sentences}$
Economic	N > 1	0.44	3220
Social	N > 1	0.38	4336
Economic	N > 25	0.81	132
Social	N > 25	0.63	104

Figure A2. Sentence-level correlation between GPT-3 ideology classification (one-shot, 175 billion parameters) and crowd-coders, Benoit et al. (2016) manifesto coding replication



B Appendix B: Non-Preregistered Estimates from GPT-4

In Figures B1 and B2, we replicate the Twitter sentiment and manifesto ideology applications from the main text using the protocol described in Le Mens and Gallego (2023). Rather than prompting the LLM for a discrete classification and then constructing a continuous measure from the resulting probability distribution (as in our pre-registered design), this approach directly prompts GPT-4 (zero-shot) for a continuous measure between 0 to 100. We generate a score for each document by taking the probability-weighted average score returned by the model. For the sentiment analysis task, each prompt includes the following instructions:

***Masterpiece Cakeshop* Prompt:**

Read this tweet posted the day after the US Supreme Court ruled in favor of a baker who refused to bake a wedding cake for a same-sex couple. What is the sentiment of this tweet? Provide your response as a score between 0 and 100 where 0 means ‘Extremely Negative’ and 100 means ‘Extremely Positive’. Respond only with this number.

***Mazars* Prompt:**

Read this tweet posted the day after the US Supreme Court ruled that sitting presidents are not immune to state criminal subpoenas, and that President Trump was obliged to disclose his tax returns to the Manhattan District Attorney. What is the sentiment of this tweet? Provide your response as a score between 0 and 100 where 0 means ‘Extremely Negative’ and 100 means ‘Extremely Positive’. Respond only with this number.

For the manifesto ideology task, each prompt includes the following instructions:

Economic Policy Prompt:

You will be provided with a text from a party manifesto. Where does this text stand on the ‘left’ to ‘right’ wing scale, in terms of economic policy? Provide your response as a score between 0 and 100 where 0 means ‘Extremely left’ and 100 means ‘Extremely right’. If the text does not refer to economic policy, return "NA". Respond *only* with your score.

Social Policy Prompt:

You will be provided with a text from a party manifesto. Where does this text stand on the ‘liberal’ to ‘conservative’ scale, in terms of social policy? Provide your response as a score between 0 and 100 where 0 means ‘Extremely liberal’ and 100 means ‘Extremely conservative’. If the text does not refer to social policy, return "NA". Respond *only* with your score.

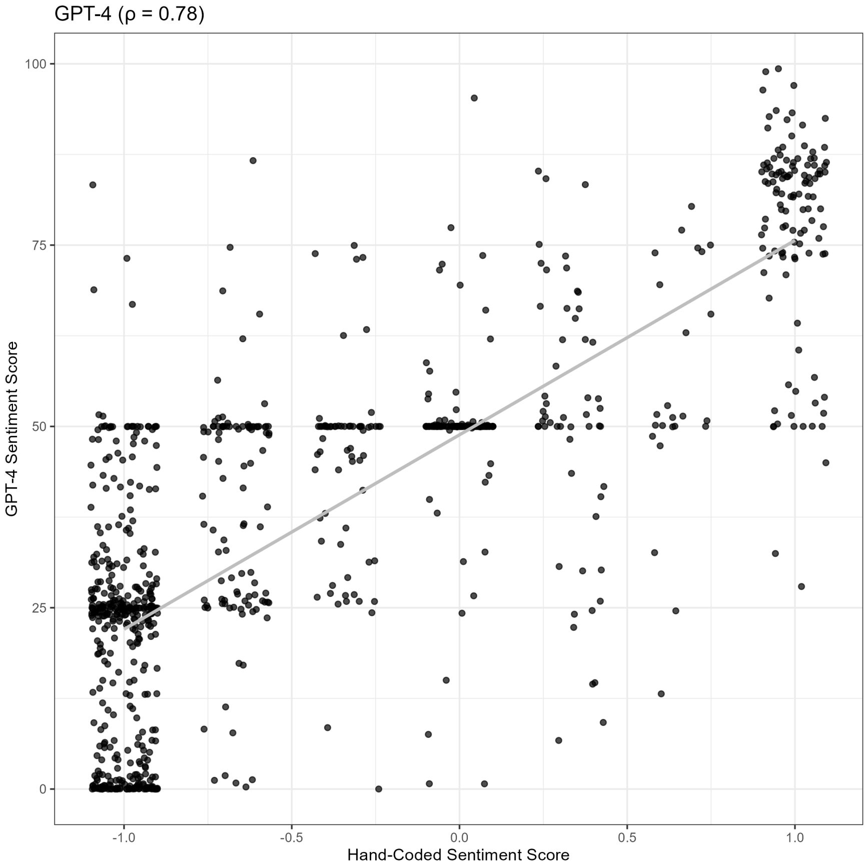


Figure B1: Replicating the tweet sentiment classifications from Figure 1

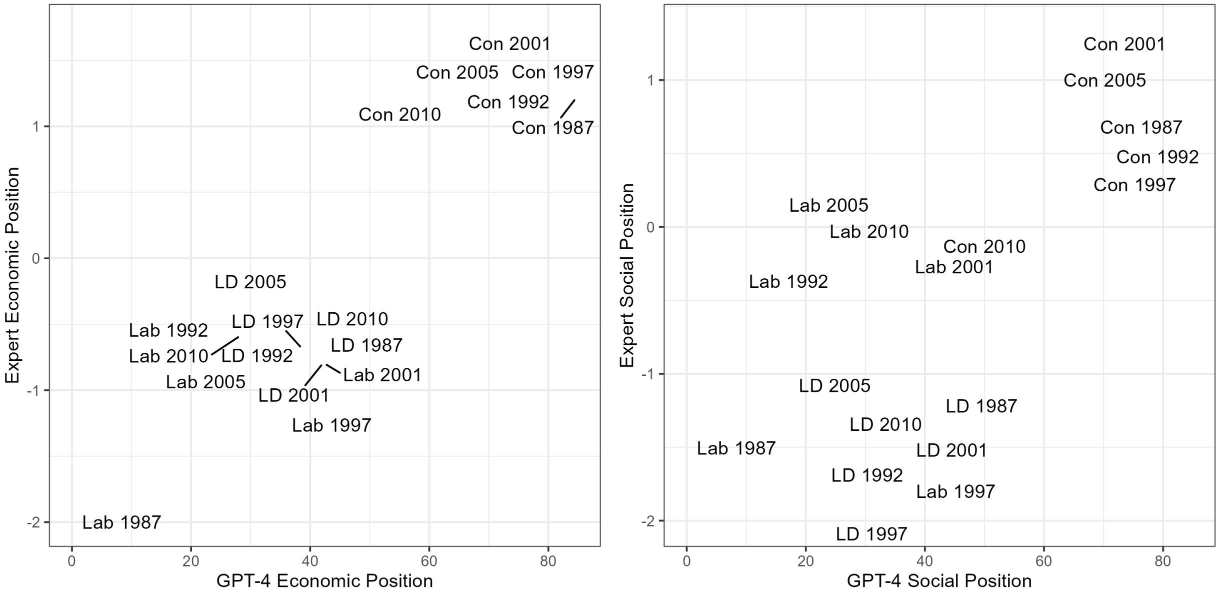


Figure B2: Replicating the manifesto ideology scaling application from Figure **??**

C Appendix C: Alternative Methods of Sentiment Classification

For the Twitter sentiment analysis task, we apply three variations of dictionary-based classifiers: BING (Hu & Liu 2004), a customized BING model, and Sentimentr (Jockers 2017). BING and Sentimentr represent different classes of dictionary-based classifiers that use adjacency matching with a pre-defined list of positive and negative terms to derive sentiment associations. At its core, the BING lexicon tokenizes text strings and derives a sentiment classification for each word based on a pre-defined lexicon. By comparison, Sentimentr classifications build on the BING-style framework by further considering the possibility of inversion rhetoric. That is, while BING associates individual words as positive or negative without any concern for their placement or usage, Sentimentr's added parameters allow it to consider conditional adverb qualifiers that might negate subsequent or preceding verbs or adjectives. For example, consider the following string: "*The Supreme Court's recent decisions are not good.*" A reliable classifier would be able to discern the negative sentiment. However, using BING and Sentimentr reveal divergent classifications that reflect their underlying parameters. Whereas Sentimentr accurately classified the string as negative (-0.08), BING actually returned a positive score (+2). BING merely observed the words *supreme* and *good* as positive qualifiers, while Sentimentr observed the inversion qualifier "not" as an indication that the sentiment was actually negative. To help reinforce BING's classifications, we included a separate classification model that removed certain terms that might increase the propensity for misclassifications. We specifically removed *supreme*, *court*, *trump*, *masterpiece*, *judge*, and *pride,* all of which were terms that frequently appeared in the Court-related tweets but could be interpreted as adjectives or verbs promoting positive or negative sentiments when they are actually being used as nouns.

D Appendix D: Lists of Topic Labels

The following table lists the most frequent topic labels returned by GPT-3 in the topic modeling application, grouped by party.

Virtue	Democratic	Republican
dedication	2472	1724
hard work	1781	1441
commitment	1671	1281
service	1605	1329
leadership	1148	1138
determination	687	407
community	567	359
excellence	553	505
perseverance	538	292
success	534	805
courage	528	470
patriotism	481	579
compassion	407	198
public service	400	214
achievement	371	272
charity	344	210
bravery	298	318
community service	295	188
justice	283	137
innovation	258	168
education	252	138
family	192	182
sacrifice	167	199
community involvement	134	352

References

Le Mens, Gaël and Aina Gallego. 2023. "Scaling Political Texts with ChatGPT." *arXiv* (arXiv:2311.16639).