Argyle et al. Study 2 Replication

1 Methodological Adjustments

To ensure that the author's token probability approach works for GPT-4o-mini and GPT-4, I slightly modified the query prompt to instruct the models to output tokens related to the candidates' names as the first and second place tokens:

Original: In the 2012 presidential election, Mitt Romney is the Republican candidate, and Barack Obama is the Democratic candidate. I voted for

Edited: In the 2012 presidential election, Mitt Romney is the Republican candidate, and Barack Obama is the Democratic candidate. Answering only with the candidate's name, I voted for"

The same changes are made for the 2016 and 2020 years. Implementing these changes ensured that the newer models behaved the same as GPT-3 when prompted.

In addition, "rom" is added to the 2012 election and "h", "hil", "cl", and "don" are added to the 2016 election as tokens for probability calculation. Certain models tend to output these incomplete words as their vote prediction for the two election years.

Finally, whereas the original authors only included tokens for probability calculation if they started with a space (" Clinton" only), I included tokens that both started with and without a space (both " Clinton" and "Clinton").

2 Proportion of Agreement

We first compare the proportion of Republican votes predicted by each model versus the actual outcome for each election year.

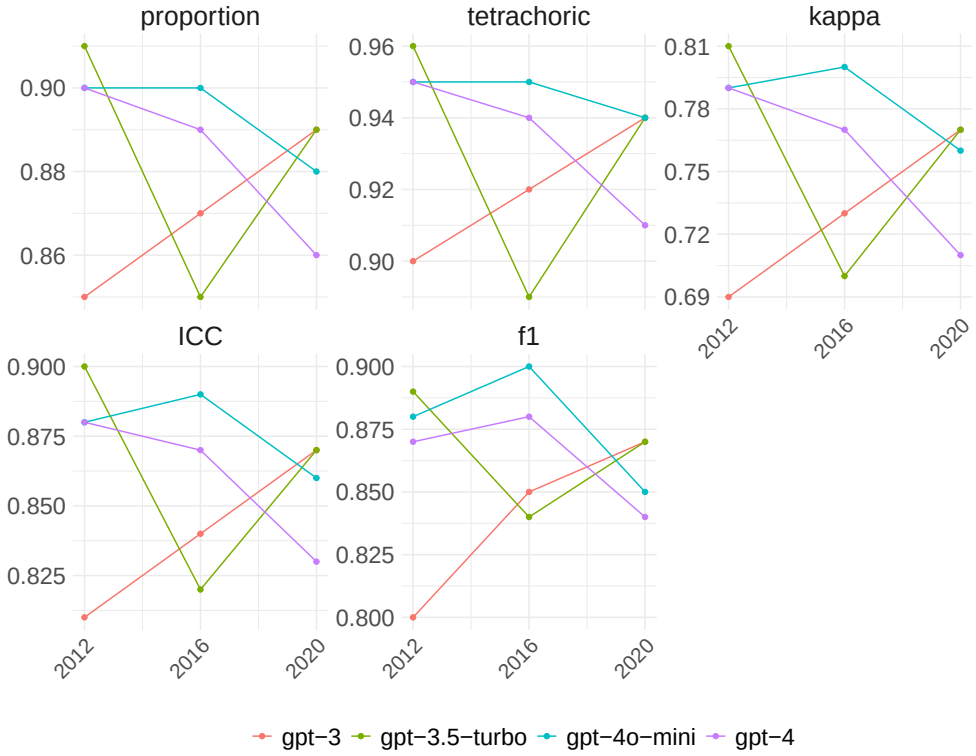
Table 1: Proportion of agreement between actual and GPT-predicted votes

Year	Model	Actual R vote	GPT-predicted R vote	P value
2012	gpt-3	0.404	0.337	0.000
2012	**gpt-3.5-turbo**	**0.404**	**0.410**	**0.563**
2012	**gpt-4o-mini**	**0.404**	**0.414**	**0.362**
2012	**gpt-4**	**0.404**	**0.394**	**0.379**
2016	gpt-3	0.477	0.419	0.000
2016	**gpt-3.5-turbo**	**0.477**	**0.466**	**0.421**
2016	**gpt-4o-mini**	**0.477**	**0.484**	**0.618**
2016	**gpt-4**	**0.477**	**0.490**	**0.375**
2020	**gpt-3**	**0.412**	**0.433**	**0.057**
2020	gpt-3.5-turbo	0.412	0.459	0.000
2020	gpt-4o-mini	0.412	0.364	0.000
2020	gpt-4	0.412	0.467	0.000

We observe that the proportion of agreement is higher for the more advanced models for the 2012 and 2016 elections, but they generally overpredict support for the Republican candidate in the 2020 election.

3 Model Performance Across Years for the 5 Metrics of Agreement

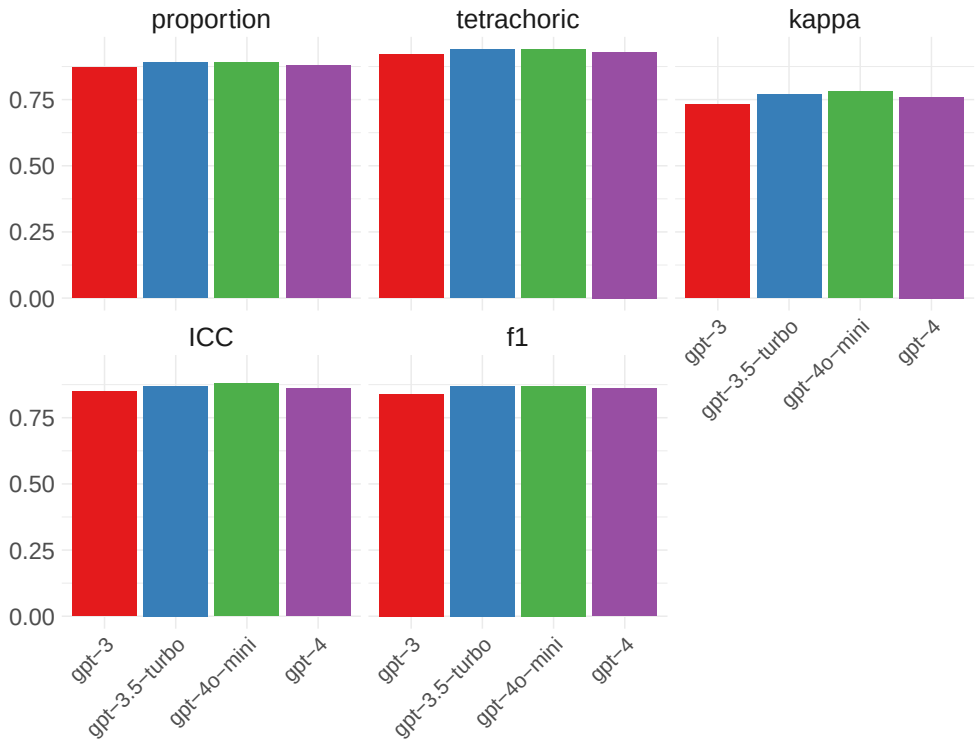
We next examine how the models perform on raw proportion agreement, tetrachoric correlation, Cohen's kappa, ICC, and f1 across years.



The three advanced models outperformed GPT-3 for the 2012 election on all five measures, while GPT-4o-mini and GPT-4 consistently outperformed GPT-3 in the 2016 election. GPT-3 and GPT-3.5-turbo achieved the same performance for the 2020 election and outperformed the other two models.

4 Average Model Performance

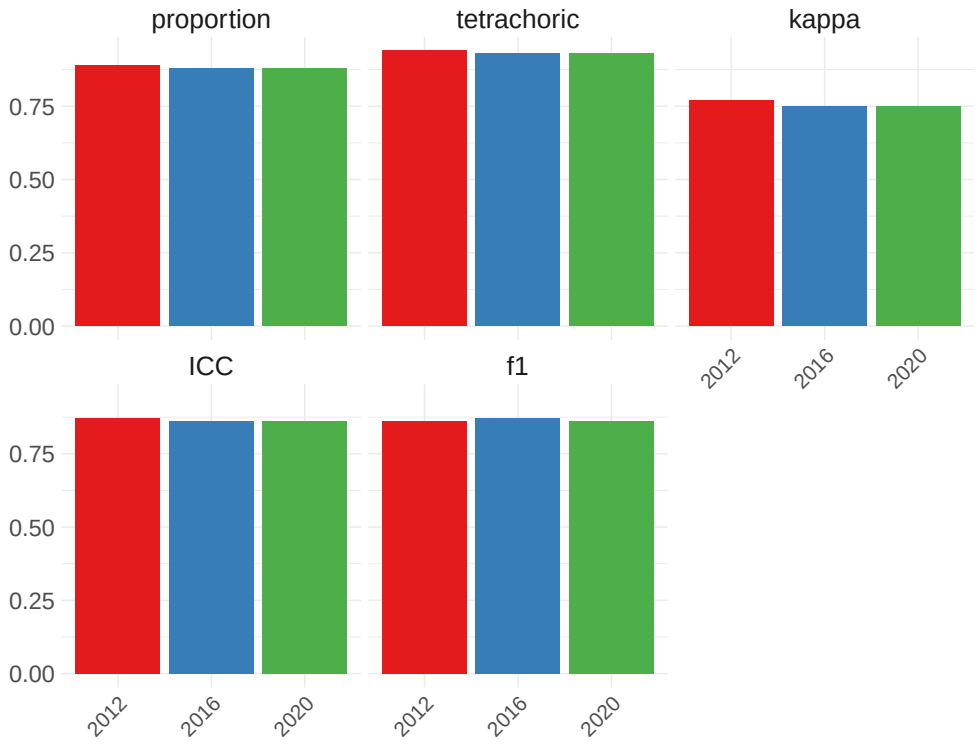
We next compare how each model performed on average across all three elections



The advanced models achieved marginally higher performance across all metrics.

5 Agreement Metrics for Each Election

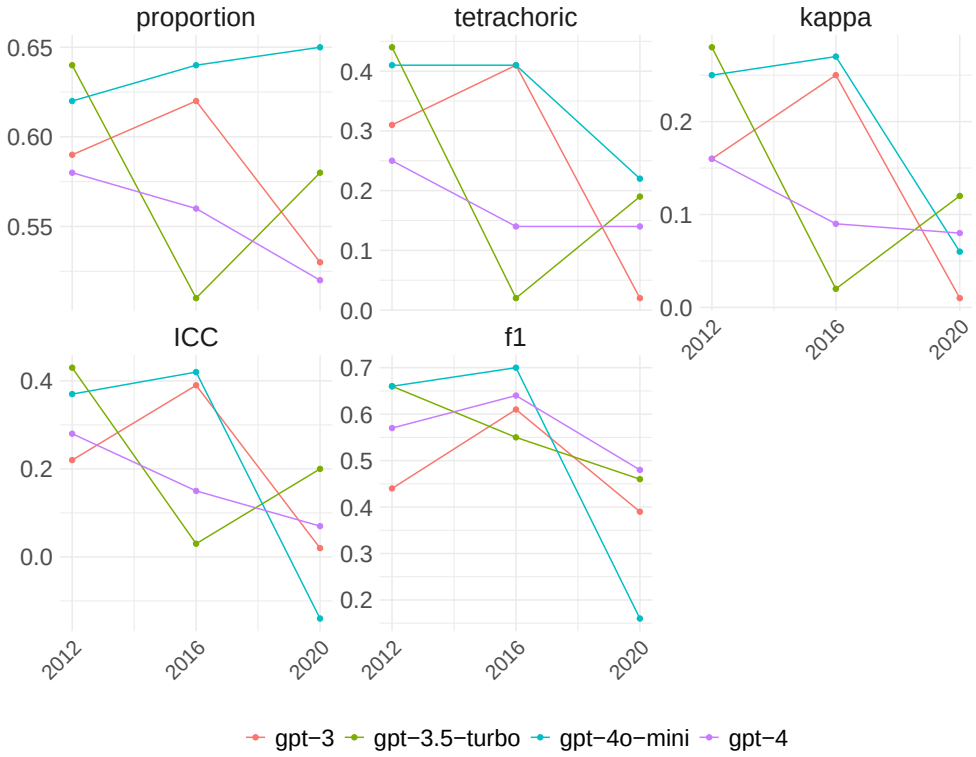
We now examine if each election year is similarly predictable in the 5 metrics by averaging the performance of all four models for each year.



All years exhibit similar predictability, although Cohen's Kappa is notably lower than the other metrics for all years.

6 Predicting Independent Voters' Outcome

The authors mentioned that the voting behavior of independents is especially difficult to predict. Let's examine if the more advanced models can do a better job.



GPT-4o-mini consistently matches or outperforms GPT-3 across all metrics for the 2012 and 2016 elections, while GPT-3.5-turbo outperforms the baseline in the 2012 and 2020 elections. GPT-4 underperforms for most metrics.