# Argyle et al. Study 3 Replication

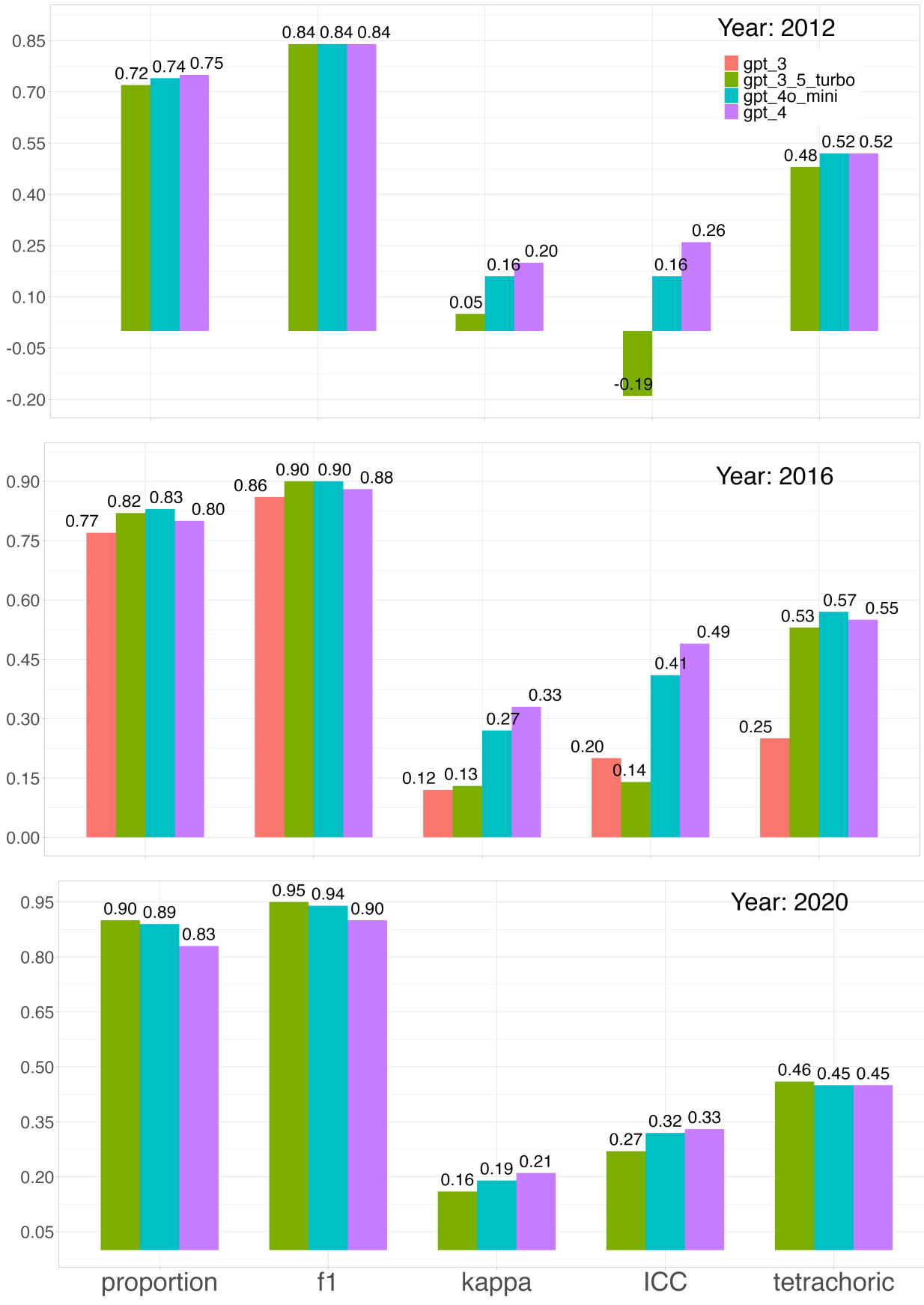
## 1 Metrics of Agreement

We compare GPT-predicted responses for political interests, political discussions, and church goer with the respondent's original response across the five metrics. It is worth noting that tetrachoric correlation is not computed for political interests as it is an ordinal variable with four levels. Also, certain models did not capture the full range of political interest categories in their predictions (for example, not a single respondent was predicted to have "very strong" political interests). The F1 score is not computed for these models.

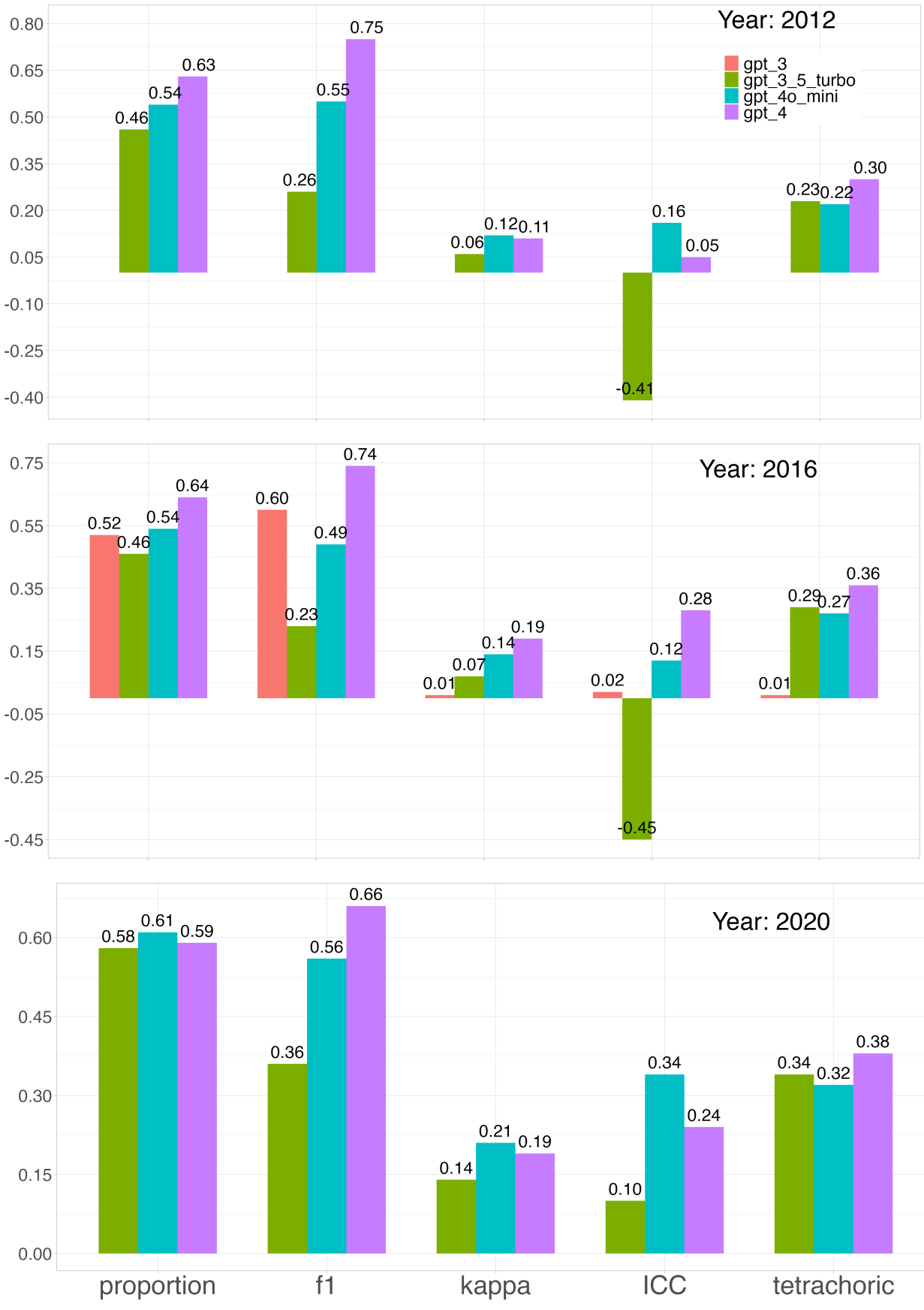
All metrics below are computed using observations with both a valid ANES response and a valid GPT prediction. As an additional exercise, I also restrict the sample to observations with available ANES responses across all variables that are used to construct the interview questions. There are no substantive changes to the results except for improved ICC scores for GPT-3.5-turbo predictions. The relevant figures can be found online in the Github repository. In addition to GPT-3.5 turbo, GPT-4o- mini, and GPT-34, the authors' GPT-3 results for 2016 are also incorporated for comparison.

Political Interests

Discuss Politics



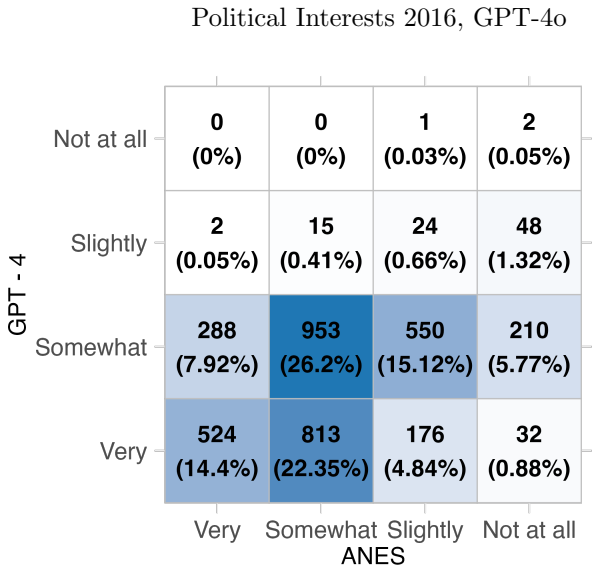
# Church Goer

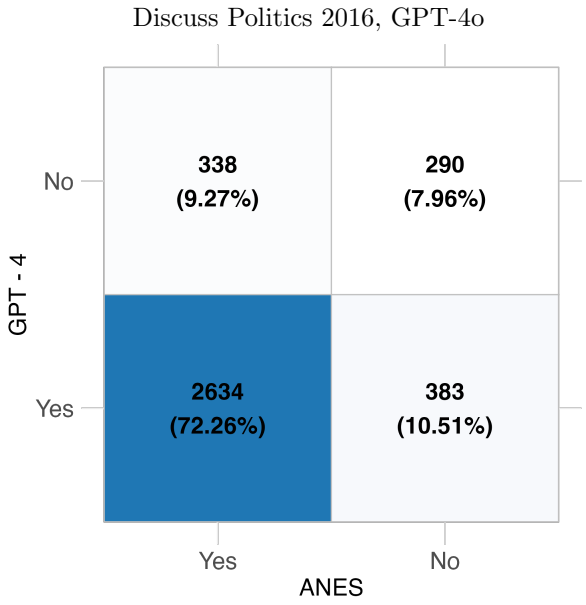

Year: 2012
gpt_3
gpt_3_5_turbo
gpt_4o_mini
gpt_4
0.80
0.65
0.50
0.35
0.20
0.05
-0.10
-0.25
-0.40
0.46
0.54
0.63
0.26
0.55
0.75
0.06
0.12
0.11
-0.41
0.16
0.05
0.23
0.22
0.30
Year: 2016
0.75
0.55
0.35
0.15
-0.05
-0.25
-0.45
0.52
0.46
0.54
0.64
0.60
0.23
0.49
0.74
0.01
0.07
0.14
0.19
0.02
-0.45
0.12
0.28
0.01
0.29
0.27
0.36
Year: 2020
0.60
0.45
0.30
0.15
0.00
0.58
0.61
0.59
0.36
0.56
0.66
0.14
0.21
0.19
0.10
0.34
0.24
0.34
0.32
0.38
proportion
f1
kappa
ICC
tetrachoric

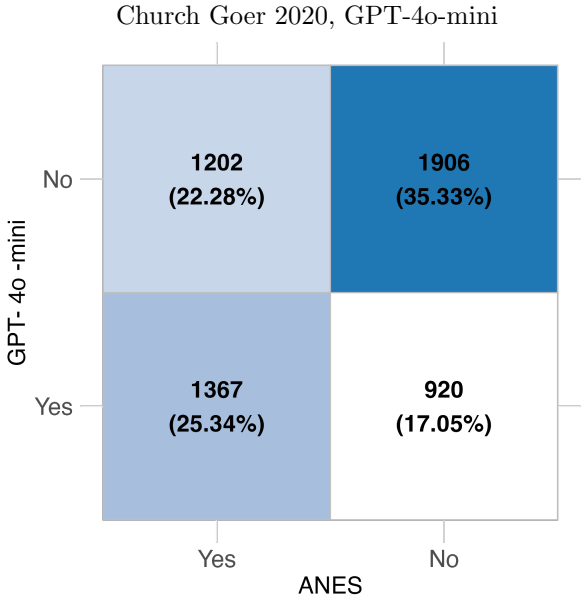
In particular, all models registered higher performances in raw proportions and F1 scores than in Cohen's Kappa and ICC, with tetrachoric correlation in between. This discrepancy suggests that the models may be registering better performance when predicting frequent classes than infrequent ones. Performance for political interests is worse because it is an ordinal variable with four classes as opposed to discuss politics' and church goer's two. It also appears that GPT-4o mini and GPT-4o outperformed the older models in most years for all target variables.

## 2 Performance Across Classes

We next plot confusion matrices for the target variables to examine prediction performance across classes. For each target variable, we choose the model-year with the highest ICC and Cohen's Kappa scores for examination.



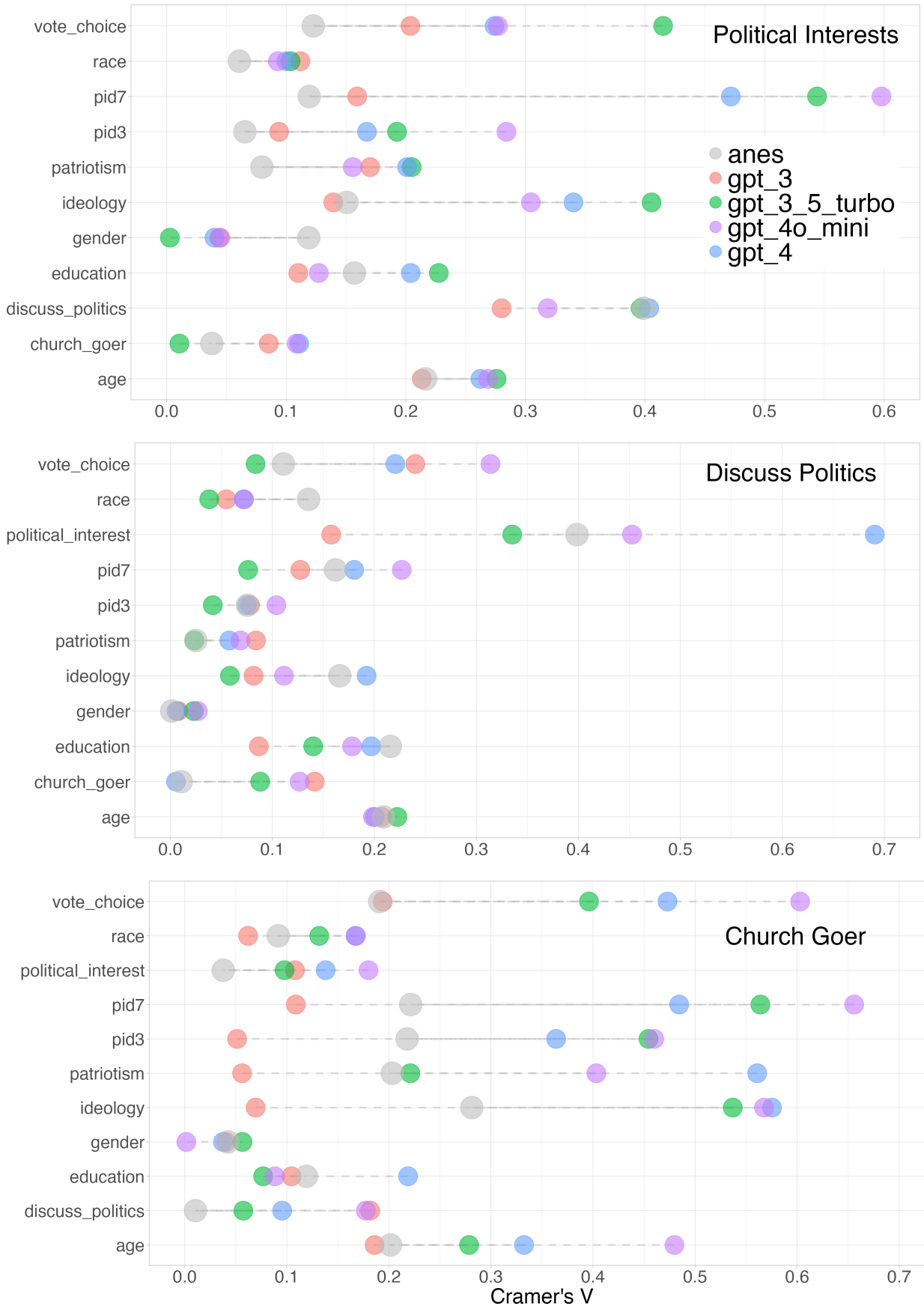



We observe that model performance is relatively unbalanced for political interests and discuss politics. The model rarely predicts the "not at all" and "slightly" classes despite that they consist of 28% of the 2016 sample. The model also falls short when it comes to predicting when respondents do not discuss politics with friends and family. These results suggest that the model may have a potential bias toward assuming that individuals are politically active. Performance is much more balanced when it comes to predicting church goers.

# 3 Cramer's V Correlations in ANES Responses vs. GPT Predictions

The authors assess whether the association between GPT-predicted responses and other ANES variables reflects the patterns observed among actual ANES variables by computing Cramer's V for each pair of variables. The same exercise is replicated below for the predictions of all three target variables from all models in 2016. The authors' GPT-3 results are included for comparison. To ensure comparability, samples are restricted to those the authors used to compute Cramer's V for GPT-3. A model demonstrates higher algorithmic fidelity when its predicted Cramér's V closely aligns with the ANES benchmark represented by the large grey dot.

Cramer's V for 2016 Predictions



Interestingly, while GPT-3 consistently obtained the worst predictive performance as seen in Section 1, its predictions appear to demonstrate more representative associations with true ANES responses compared to the newer models for political interests and church goer. In the authors' own words, this shows that while GPT-3 predictions show worse correspondence with ANES data at the individual level for these three target variables, its overall distribution better matches the overall distribution of the real data.