Learning from Convenience Samples: Fine-Tuning LLMs to Predict Voting Choices from Biased Survey Samples

Tobias Holtdirk[1,2] **Dennis Assenmacher**[3] **Arnim Bleier**[3] **Claudia Wagner**[3,4,5]

[1]LMU Munich [2]Munich Center for Machine Learning (MCML)
[3]GESIS – Leibniz Institute for the Social Sciences
[4]RWTH Aachen University [5]Complexity Science Hub Vienna

tobias.holtdirk@lmu.de
{dennis.assenmacher,arnim.bleier,claudia.wagner}@gesis.org

Abstract

Survey research increasingly relies on convenience samples as rising costs and nonresponse make representative probability sampling more difficult. This creates a central challenge for public opinion research and electoral studies: how to recover reliable population-level inferences from socially and demographically biased data. While recent work has explored the use of large language models (LLMs) for simulating survey responses, most studies focus on the U.S. population and assume relatively favourable data conditions with limited or random missingness. In this work, we investigate whether LLMs can leverage contextual knowledge about socio-demographic and political patterns to predict vote choice in the German population under sampling constraints. We systematically vary both the size and composition of training data, ranging from large representative samples to small, biased convenience samples, and compare zero-shot prompting and supervised fine-tuning against established tabular classifiers such as CatBoost.

Our results show that when representative training data are available, fine-tuned LLMs achieve predictive performance comparable to strong tabular baselines and substantially outperform zero-shot prompting. Under small and biased convenience samples, however, fine-tuned LLMs more accurately recover both individual-level vote choice and population-level distributions than zero-shot approaches and often outperform tabular methods. These findings suggest that fine-tuned LLMs can combine sample-specific information with contextual knowledge acquired during pre-training, enabling more robust inference under conditions of sampling bias and missing data.

1 Introduction

For decades, the empirical social sciences have relied on probability-based survey samples as one of their main data sources. However, survey-based research is witnessing transformative changes, including rising costs, declining participation, and new modes of participation (Berinsky, 2017; Couper, 2017). These challenges particularly affect surveys on political opinions and voting behaviour, where systematic nonresponse biases can impair election forecasting (Feldman and Mendez, 2022; Clinton et al., 2022). As representative sampling becomes more difficult and costly, researchers increasingly turn to *convenience samples*—non-probability samples in which participants are selected because they are easy to access or inexpensive to recruit.

Predicting population behaviour or responses from convenience samples can be understood as a form of distribution shift, where the observed sample on which prediction models are trained is not representative of the target population that is present in the test data. In this context, large language models (LLMs) offer an interesting alternative to traditional machine learning methods because they can draw not only on the biased training data of the convenience sample but also on patterns learned during large-scale pre-training. In principle, exposure to broad pre-training data could compensate for subpopulations that are underrepresented in the convenience sample.

Whether this potential is realised, however, depends on how LLMs are used. When simply prompted zero-shot, LLM responses can be biased and lack variation (Bisbee et al., 2024), often displaying left-leaning tendencies in political orientation (Santurkar et al., 2023; von der Heyde et al., 2025). Zero-shot LLMs also face technical issues such as first-token probabilities not matching text responses (Wang et al., 2024) and prompt sensitivity (Bisbee et al., 2024), suggesting zero-shot prompting alone may be insufficient for survey prediction. Advances in the efficiency of fine-tuning, such as quantisation (Dettmers et al., 2023) and parameter-efficient adaptation (Hu et al., 2022; Liu

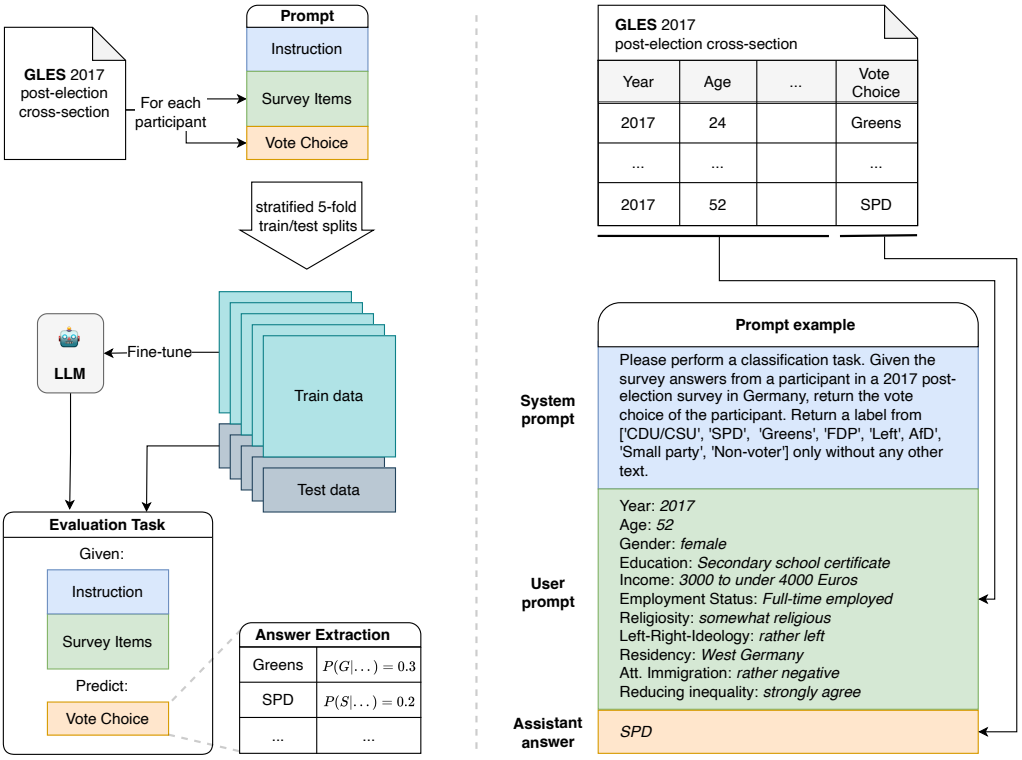

Figure 1: **Approach to simulate survey answers with LLMs.** We prompt and fine-tune LLMs to extract their answer for the question "Which party did you vote for in the federal election?". We compare LLM-generated answers with real survey answers. The prompt design consists of three parts: the system prompt, the user description, and the answer. For fine-tuning, we provide the real vote-choice labels of training participants as assistant completions in the depicted format. For the evaluation, we only present the system prompt and the user prompt to the model and require the model to generate the answer.

et al., 2022), together with strong open-source models (Grattafiori et al., 2024), make fine-tuning a more accessible alternative that can combine the pre-trained knowledge with the available sample data, addressing these issues while allowing local deployment for confidential survey data.

In this paper, we predict the voting behaviour of respondents in the German Longitudinal Election Study (GLES) (GLES, 2020) with LLMs (see Figure 1) and address two research questions:

> **RQ1:** Are LLMs more accurate and robust than tabular classifiers when **large, random samples** are available for training/fine-tuning?

> **RQ2:** Are LLMs more accurate and robust than tabular classifiers when **only small, biased convenience samples** are available for training/fine-tuning?

We compare LLMs (zero-shot and fine-tuned) against tabular classifiers (random forest, CatBoost, logistic regression) across four scenarios (Figure 2). All experiments predict the vote choice of the same 20% randomly selected GLES participants but vary the training set drawn from the remaining 80%: **E1** uses the full 80%; **E2** restricts training to students; **E3** to participants from Thuringia, an East German state whose 2017 voting pattern deviated markedly from the national average; **E4** to unemployed participants. E2–E4 test to what extent LLMs can learn from small, biased convenience samples and generalise to the broader population.

Our results show that when large random samples are available (E1), fine-tuned LLMs match tabular classifiers, outperform zero-shot LLMs, and remain robust when key predictors are removed. When only small, biased convenience samples are available (E2, E3), fine-tuning small open-source LLMs (3B–8B) generally yields more accurate individual-level and aggregate predictions than zero-shot LLMs or tabular baselines, because these models combine sample-specific patterns with broad contextual knowledge acquired during pre-training.

2 Related Work

Zero-shot opinion prediction. Most prior work uses zero-shot prompting, typically on OpenAI models. Argyle et al. (2023) reported that GPT-3, conditioned on socio-demographic backstories from the American National Election Studies (ANES), closely reproduces vote-choice patterns in the aggregate and across most demographic subgroups, though they deliberately do not evaluate individual-level correspondence. Later work paints a less optimistic picture: Santurkar et al. (2023) showed systematic biases toward certain demographic groups and political orientations on the American Trends Panel; Dominguez-Olmedo et al. (2024) showed on American Community Survey (ACS) data that zero-shot survey responses are governed by ordering and labelling biases and become near-uniform once these biases are adjusted for; and Bisbee et al. (2024) concluded that current LLMs cannot replace human ANES respondents. Outside the U.S., von der Heyde et al. (2025) and Qu and Wang (2024) report lower performance for German and other non-U.S. populations.

Fine-tuning for opinion prediction. Several studies fine-tune or adapt open-weight language models on domain-specific data: Jiang et al. (2022) fine-tune GPT-2 on partisan Twitter communities, Chu et al. (2023) adapt BERT to US media-outlet content, and Ahnert et al. (2025) fine-tune LLMs on Twitter time slices. Recent work trains on closed-ended response distributions rather than individual responses (Cao et al., 2025; Suh et al., 2025). Our approach enables training on both closed- and open-ended questions and is compatible with fine-tuning APIs, providing a more general framework.

LLMs for imputation. Pre-trained language models have been fine-tuned for imputation (Mei et al., 2021), and LLMs have been applied to data-wrangling tasks including imputation more broadly (Narayan et al., 2022); the HELM benchmark (Liang et al., 2023) includes a corresponding imputation scenario. Existing work largely focuses on data missing completely at random. An exception is Kim and Lee (2023), who train a dense network on LLM embeddings for U.S. surveys. In contrast, we directly fine-tune instruction-tuned LLMs and evaluate on German data, which is known to be more challenging (von der Heyde et al., 2025).

3 Methods and Experiments

We predict the target variable "vote choice" using survey variables that the literature on voting behaviour in Germany identifies as important predictors (Schmitt-Beck et al., 2022; Schoen et al., 2017; Klein, 2014); the selected items are detailed in Section 3.1. We compare fine-tuned LLMs against zero-shot LLMs and tabular classifiers (Figure 1) with stratified 5-fold cross-validation, reporting means and 95% confidence intervals. We evaluate individual-level classification with macro-F1 and distributional accuracy with aggregated vote shares and total variation distance (TVD).

3.1 Survey Data Processing

We use the German Longitudinal Election Study (GLES) Cross-Section Cumulation 2009–2017 (GLES, 2020), which contains pre- and post-election face-to-face interviews. We predict vote choices in the 2017 post-election cross-section, enabling comparison with von der Heyde et al. (2025).

Item selection and mapping. Survey items were selected based on factors associated with vote choice in German elections (Schmitt-Beck et al., 2022; Schoen et al., 2017; Klein, 2014): demographics (age, gender, education, income, employment status), religiosity, left–right ideology, party identification (and its strength), place of residence, and attitudes toward immigration and income inequality. Each item maps one-to-one to a feature, and the participant's self-reported vote choice is the target. We drop the infrequent categories *invalid vote* and *don't know*, and keep the original German wording since translation could obscure cultural nuance.

Missing values. Following von der Heyde et al. (2025), we drop respondents with missing vote choices. However, we *retain* other missing values rather than imputing them, since the type of missingness (e.g., *no answer*, *interview aborted*, or higher missingness for *income*) may itself be informative.

3.2 LLM Setup

Prompt design. Our task is instruction-based, so we use instruction-tuned chat models. Each instruction consists of a system prompt defining the task and a user prompt filled with a participant's survey responses (Figure 1); the assistant response

is the predicted vote choice. During fine-tuning we supervise on the assistant response; at evaluation only the system and user prompts are presented.

Model selection. For robustness, we test a range of model sizes across two families. We use three Llama-3 instruction-tuned sizes: Llama-3.2-1B and Llama-3.2-3B (Meta AI, 2024), and Llama-3.1-8B (Grattafiori et al., 2024) (we omit the "Instruct" suffix below). We additionally run Qwen2.5 models from 0.5B to 7B (Qwen et al., 2025) in Appendix B.

Supervised fine-tuning. We fine-tune on a 40 GB A100 partition in 16-bit precision using LoRA (Hu et al., 2022) with rank $r = 256$ applied to all linear layers. We use rank-stabilised scaling ($\alpha/\sqrt{r}$ instead of α/r, with scaling factor $\alpha = 8$) (Kalajdzievski, 2023), which was proposed precisely to keep learning effective at such high ranks. Training uses batch size 1, which fits the memory budget of the GPU partition, and runs for three epochs; all other hyperparameters remain at their library defaults. Training is conducted on completions only, where the loss is computed only on the assistant response tokens and not on the system and user prompt tokens. Formally, given a sequence of tokens $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ where tokens x_1 to x_k represent the prompt and tokens x_{k+1} to x_n represent the completion, the loss function is:

$$\mathcal{L} = -\sum_{i=k+1}^{n} \log P(x_i | x_1, \ldots, x_{i-1})$$

This ensures that the model learns to generate appropriate completions without being penalised for the fixed prompt content.[1]

Answer extraction and probabilities. We follow Argyle et al. (2023) and take the highest-probability token at the first answer position; the eight vote-choice categories (CDU/CSU, SPD, Greens, FDP, Die Linke, AfD, small party, non-voter) each have a distinct initial token, so this is unambiguous. We obtain a probability distribution by applying softmax only over the logits of these eight initial tokens. First-token probabilities are known to be unreliable in zero-shot settings: they need not match the model's actual text answer (Wang et al., 2024), and the elicited answer distributions are confounded by ordering and labelling biases (Dominguez-Olmedo et al., 2024). After fine-tuning, however, the mass on tokens outside the eight answer categories becomes negligible.

[1] https://huggingface.co/docs/trl/v0.13.0/en/sft_trainer

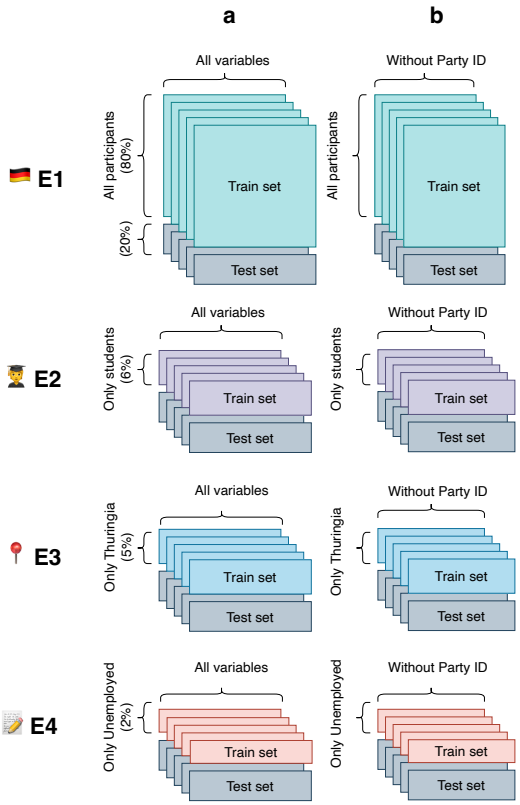

Figure 2: **Experiments.** Each experiment uses a different subset of the GLES data, reduced either by removing survey variables (columns) or participants (rows).

3.3 Experiments

All experiments use the same stratified 5-fold cross-validation splits, where each fold holds out 20% of all GLES respondents as the test set. The training set is drawn from the remaining 80% but varies by experiment: **E1** uses the full 80%, simulating the scenario where responses are missing completely at random (RQ1). **E2–E4** restrict the training set to biased convenience samples[2] within that 80%, simulating systematic missingness (RQ2):

> **E2:** We restrict the training set to GLES participants whose reported occupation is "student" (enrolled in school or university), which amounts to 6% of all participants. This simulates a scenario in which a researcher can

[2] We use "convenience sample" as shorthand for biased subsamples constructed from the GLES data to mimic convenience sampling. These simulate the effect of training on systematically restricted respondent groups, but are not *real* convenience samples (e.g., respondents self-selecting into a non-probability survey).

only recruit students—a common convenience sample in the social sciences (Mullinix et al., 2015)—and must generalise from this subpopulation to the full electorate.

E3: We restrict the training set to GLES participants residing in the state of Thuringia (5% of participants). This simulates a regionally limited survey. As a former East German state, Thuringia had a distinctly different 2017 voting profile from the national average; for example, AfD's second-vote share was substantially higher in Thuringia than nationally (Die Bundeswahlleiterin, 2017). This makes it a useful regionally biased training subsample.

E4: We restrict the training set to GLES participants who reported being unemployed (2% of participants). This is the smallest and most biased convenience sample we test, motivated by evidence that unemployed individuals are more likely to participate in surveys when monetary incentives are offered (Singer and Kulka, 2002).

These experiments pose a domain generalisation problem: restricting training to a specific subpopulation changes not only the covariate distribution $P_{\text{train}}(X) \neq P_{\text{test}}(X)$, but likely also the conditional $P(Y \mid X)$, since subpopulation membership may correlate with factors not included in our feature set (e.g., social milieu, peer effects) that also influence vote choice.

We additionally run each experiment in two variants: variant **a** uses all survey variables, while variant **b** removes "party identification", which is by far the strongest predictor of vote choice in our feature set (see Appendix C) and tends to be missing whenever vote choice is also missing. This makes the "b" variants more realistic for practical settings.

Baselines. We compare against three **tabular classifiers** (logistic regression, random forest, CatBoost), **zero-shot LLMs** (GPT-4o, 2024-11-20, and our open-source models without fine-tuning), and a **prior-matched random classifier** that samples from the training label distribution. We also evaluate **importance-weighted variants** of the tabular classifiers to correct for covariate shift in the convenience sample experiments (Table 1).

Evaluation metrics. We are interested in assessing the predictive performance for individual respondents, as well as the aggregated vote share of the population. We use the following metrics:

- **Macro-F1 score**: The macro-F1 score of the predicted vote choices, calculated as $\text{F1}_{\text{macro}} = \frac{1}{k}\sum_{i=1}^{k} \text{F1}_i$, where F1_i is the F1 score for party i and k is the number of vote choice categories.
- **Aggregated vote share**: We calculate the aggregated vote share of the predicted vote choices as $\hat{P}(i) = \frac{1}{N}\sum_{j=1}^{N} \hat{p}_j(i)$, where N is the number of test set participants and $\hat{p}_j(i)$ is the predicted probability that participant j votes for party i. The true aggregated vote share is $P(i) = \frac{1}{N}\sum_{j=1}^{N} y_j(i)$, where $y_j(i)$ is 1 if participant j voted for party i and 0 otherwise.
- **Total variation distance (TVD)**: $\text{TVD}(\hat{P}, P) = \frac{1}{2}\sum_{i=1}^{k} |\hat{P}(i) - P(i)|$, where $\hat{P}$ and P are the predicted and true aggregated vote shares over the k vote choice categories.

Reproducibility. All experiments use fixed random seeds. GLES Cross-Section Cumulation 2009–2017 is openly available (GLES, 2020), and our code is released.[3]

4 Results

4.1 RQ1: Missing Completely At Random

With a 20% random test split, tabular classifiers and fine-tuned Llamas perform on par and both clearly outperform all zero-shot models in macro-F1, including GPT-4o (Figure 4, E1). Zero-shot rankings are also unstable: e.g., Llama-3.1-8B trails the smaller Llama-3.2-3B in E1a, whereas the fine-tuned models are consistent across sizes and families. The same picture holds for TVD (Figure 5): fine-tuned and tabular models approximate the true distribution, while zero-shot models systematically deviate. Figure 6 (E1) shows GPT-4o over-predicting parties at the extremes and rarely predicting non-voters, and zero-shot Llama-3.1-8B heavily over-predicting Die Linke. These patterns are consistent with the demographic and opinion-alignment biases documented by Santurkar et al. (2023) and with refusal behaviour, where the model avoids committing to certain answer categories (Wang et al., 2024).

[3] https://github.com/tobihol/convenience-sample-finetuning

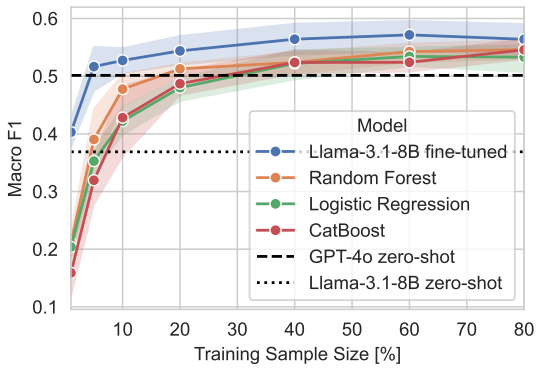

Figure 3: **Sample efficiency comparison.** We vary the size of the training sample (1%, 5%, 10%, 20%, 40%, 60%) of the E1a setup, drawn as *random* sub-samples of GLES participants. Additionally, the original 80% training sample is included. Because the sweep uses random sub-samples (not convenience-style biased ones), it isolates sample-size effects from sampling-bias effects.

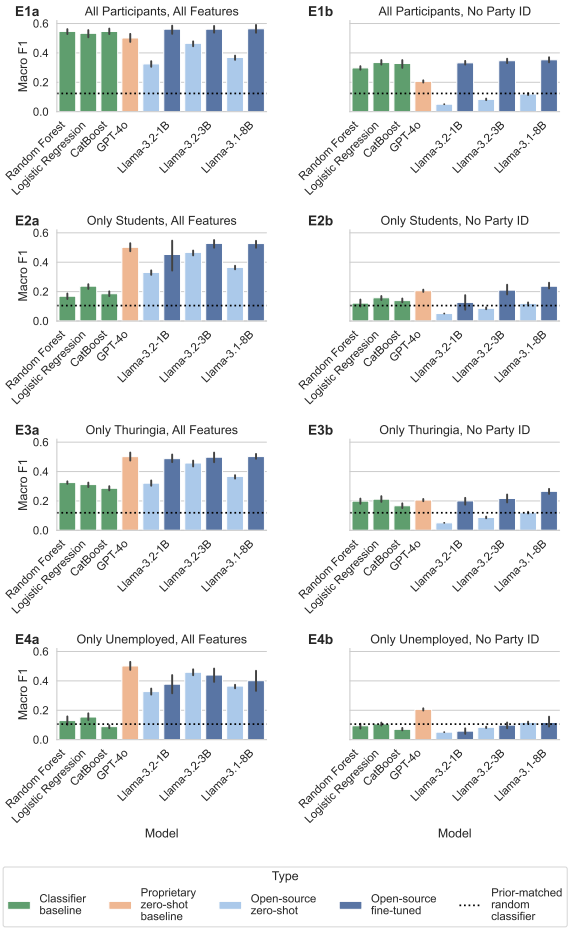

Figure 4: **Macro-F1 scores (higher is better)** of the predicted vote choices. The left column shows results using all survey variables as features. The right column shows results without using "party identification". The rows show different subsets of participants used for training.

The similar performance of all tabular classifiers and fine-tuned models in E1 suggests that when large random training samples are available, fine-tuning LLMs offers no clear advantage over traditional classification methods, which require less compute. The sample-size sweep in Figure 3 refines this picture, however: zero-shot LLMs are only attractive below 5% training data; from roughly 5% up to about 40% the fine-tuned Llama clearly outperforms all tabular classifiers; only as the sample approaches the full 80% do tabular methods catch up. Because the sweep varies sample size on random sub-samples (not convenience-style biased ones), it isolates sample-size effects from sampling-bias effects. Removing the strong predictor "party identification" (E1a→E1b) degrades all methods, but hurts zero-shot LLMs most (see gap between panels E1a and E1b in Figures 4 and 5).

4.2 RQ2: Convenience Samples

As expected, the performance of all models drops when training data are limited to biased subsets of the GLES data. However, the extent of the decrease differs across experiments. In experiments E2 and E3, fine-tuned LLMs win: the 3B and 8B fine-tuned Llamas are better at modelling the true vote share distribution than both the zero-shot LLMs and the tabular classifiers. For the smallest sample (E4, 2% of participants), the picture flips: zero-shot GPT-4o becomes the strongest model, and fine-tuned LLMs no longer outperform the tabular baselines. Applying importance weighting to correct for covariate shift in the tabular baselines yields mixed results: CatBoost benefits in some experiments, but the overall gap to fine-tuned LLMs remains (see Table 1).

E2: Only Student Data Figure 4 (E2a) shows that fine-tuned open-source LLMs and GPT-4o are better able to predict the vote choice of GLES participants in the test split, compared to tabular classifiers. Figure 5 (E2a) indicates that the 3B and 8B fine-tuned Llama models in particular achieve considerably lower total variation distance, suggesting that fine-tuning can help to reduce the deviation from the true distribution, even when training on biased data. Deviations, however, still exist. GPT-4o produces especially large deviations for non-voters and the extreme parties (Die Linke and AfD; see Figure 6). CatBoost also overpredicts non-voters and underpredicts CDU/CSU considerably.

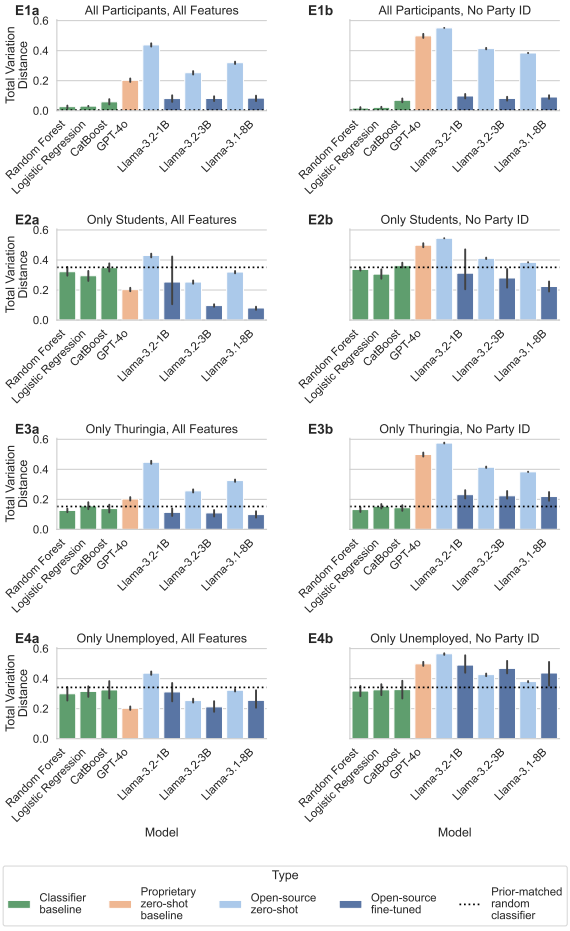

Figure 5: **Total variation distance (lower is better)** between the predicted distribution and the true distribution of vote choices. The left column shows results using all survey variables as features. The right column shows results without using "party identification". The rows show different subsets of participants used for training.

When the important predictor party identification is removed from the training data, zero-shot models in particular start to overestimate the share of Die Linke and the AfD (cf. Figure 6 E2b).

E3: Only Thuringia Region Figure 4 shows that F1 scores for E3 are higher across all models compared to E2, which is expected since the Thuringia training distribution is closer to the national vote distribution than the student sample (see Figure 6, training subset). In both variants, E3a and E3b, the 8B fine-tuned Llama model outperforms the classifier baseline with respect to the macro-F1. However, the difference in E3b is notably smaller than in E3a. Both fine-tuned LLMs and tabular classifiers decrease in performance when an important predictor variable is removed and show comparably poor performance.

Figure 5 shows that the fine-tuned LLMs and tabular classifier perform similarly with respect to the TVD. Figure 6 also shows that the fine-tuned LLMs and tabular classifier make similar mistakes (e.g., they overestimate CDU/CSU and underestimate the Greens), directly reflecting Thuringia's training distribution, which over-represents CDU/CSU and AfD relative to the national average. Zero-shot LLMs again strongly overestimate the share of Die Linke and the AfD, as in the previous experiment.

E4: Only Unemployed Data This experiment shows the lowest scores for the fitted models, which is expected given that it uses the smallest training set. The fine-tuned models fail to outperform the zero-shot setting, as reflected by their similar F1 scores and total variation distances. In this case, the zero-shot proprietary model achieves the best performance among all models, but only marginally so. In Figure 6, we can see that both fine-tuned models and CatBoost overpredict non-voters for E4, which fine-tuned models did not do for E2, although the two training sets both over-represent non-voters relative to the complete dataset distribution (see the training-subset lines in Figure 6).

This experiment shows that for extremely small convenience samples (2%), fine-tuned LLMs and tabular classifiers perform worse than zero-shot LLMs.

5 Discussion and Conclusion

The findings of this study contribute to the growing literature on the use of large language models for predicting survey responses. Our results show that the effectiveness of LLMs for predicting survey responses depends critically on the amount and composition of available training data.

When only small but biased convenience samples are available, such as datasets composed exclusively of students, fine-tuning small open-source LLMs (3B–8B) yields substantial gains: fine-tuned models recover both individual-level predictions and aggregate vote-share distributions more accurately than zero-shot LLMs, including GPT-4o, and often outperform tabular baselines. These gains likely arise because fine-tuned models can draw simultaneously on sample-specific correlations and broader knowledge that models have acquired in the pre-training phase.

However, the advantages of fine-tuning diminish when training sets become extremely small. In the E4 condition, where only 2% of participants were available for training, fine-tuned models no

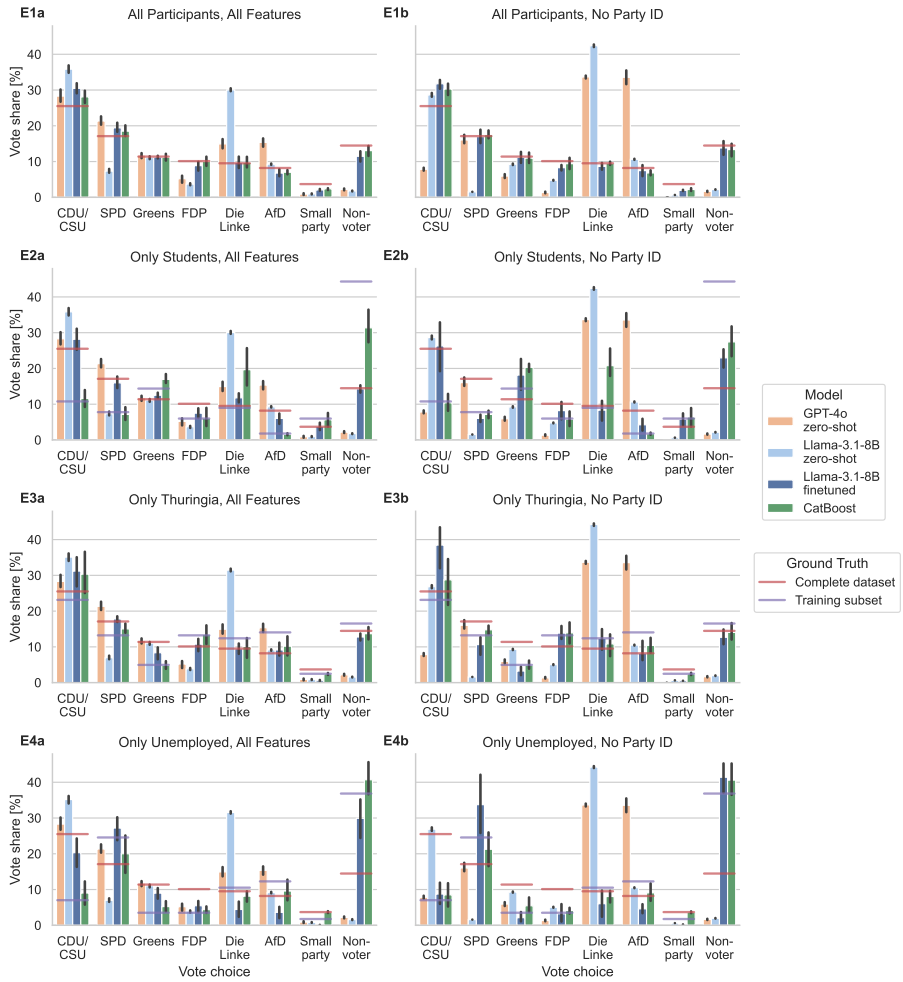

Figure 6: **Aggregated vote share** of the predicted vote choices by model across all experiments (E1–E4). The red line shows the true vote share in the complete dataset; the purple line shows the vote share within the respective biased training subset. For E1 both the tabular classifier and the fine-tuned model perform similarly well, in the case with and without having the strong predictor party identification in the data. The zero-shot models struggle with predicting non-voters in (a) and additionally overpredict more extreme parties (the AfD being the most right-wing party and Die Linke the most left-wing) in (b). For E2a, we see that fine-tuned LLMs are able to match the true distribution well. This degrades somewhat for E2b, and similarly to the CatBoost baseline, overfits to the student population by predicting more left-leaning parties and more non-voters. For E3, the training vote share distribution is more similar to the complete data. For E4, non-voters are again overrepresented in the training data, as in E2; here, however, the fine-tuned LLM does not generalise well to the overall population and overpredicts non-voters, similar to CatBoost.

longer outperformed their zero-shot counterparts, and the zero-shot proprietary model achieved the strongest results. This suggests that very small samples provide insufficient task-specific signal for effective supervised adaptation, making zero-shot inference from large pretrained models a more reliable choice.

When medium- to large-sized random samples are available (E1), fine-tuned LLMs and tabular classifiers perform similarly, and given the substantially higher computational cost of LLMs, we do not consider them promising for imputing small amounts of randomly missing data in surveys. A caveat is that zero-shot proprietary models like GPT-4o are not fully reproducible due to model updates and nondeterminism.

Our zero-shot results also speak to the concern of data contamination (see Limitations): if the models had simply memorised the widely documented 2017 election results, their zero-shot predictions should closely match the true aggregate vote shares. Instead, we observe substantial deviations in the zero-shot setting (e.g., systematic over-prediction of AfD and Die Linke, under-prediction of non-

voters; cf. Figure 6), which provides indirect evidence against strong contamination.

It is important to note that even the best-performing models often achieve macro-F1 below 0.6 in the convenience-sample settings (E2–E4). Our results should therefore be read as evidence that fine-tuned LLMs are a *promising direction* rather than a ready-to-deploy solution, especially in multi-party systems like Germany's where small parties can decisively influence coalitions.

Taken together, these findings offer practical guidance, although we caution that they are derived from a single election in a single country and from convenience samples constructed retrospectively (see Limitations): when training data are extremely limited, zero-shot prompting of a strong proprietary model remained competitive in our experiments. When small-to-medium but biased convenience samples are available, fine-tuning small open-source LLMs was the most promising approach in our setting, as these models leverage both sample-specific signals and world knowledge. With medium-to-large random samples that reflect the target population, traditional machine-learning models remain strong options.

Future work should test generalisation to other survey variables, populations, and response formats, and systematically establish the amount and type of training data required for fine-tuning to yield stable improvements.

Limitations

A potential concern is data contamination: the 2017 German federal election results are widely documented online and likely part of the pre-training corpora of the models we evaluate (Llama-3, GPT-4o). In principle, LLMs could leverage memorised aggregate statistics rather than reasoning about individual respondent profiles. While the zero-shot deviations discussed in Section 5 argue against strong contamination, we cannot fully rule out indirect advantages from pre-training exposure, and future work should evaluate on elections that post-date the models' training cutoff.

Our convenience samples are also constructed retrospectively from a probability-based survey rather than collected as real non-probability samples. This design lets us compare models against a common held-out test set and isolate the effect of biased training composition, but it may understate or overstate the difficulties of real convenience samples, where recruitment, motivation, survey mode, and measurement error can differ from probability samples.

Temporal validity poses a further concern: voting behaviour shifts across election cycles, and it is unclear how our findings transfer to later elections (e.g., 2021 or 2025). We also use zero-shot LLMs as baselines without prompt variation, acknowledging that their performance could change with alternative prompts. Finally, while we focused on fine-tuning, techniques like retrieval-augmented generation or in-context learning with long-context models are promising directions for future work.

Ethical Considerations

The data used in this study are derived from survey results available through GESIS under access conditions, with all responses anonymised. The survey data contain no personally identifying information or offensive content. An important ethical concern with simulating public opinions using LLMs is that generated responses may misrepresent genuine public sentiment. Because LLMs base their outputs on patterns found in historical and potentially biased data, they could reinforce prejudices, obscure minority perspectives, or amplify simplistic narratives. Our results suggest that fine-tuning can help reduce but not fully mitigate these issues.

Beyond misrepresentation, predicting voting behaviour itself carries risks for democratic processes. Models that infer vote choice from socio-demographic and attitudinal profiles could be misused for targeted political advertising or demobilisation campaigns aimed at groups predicted to vote a certain way, and synthetic "respondents" could be passed off as genuine public opinion to influence public debate or election coverage. Inaccurate predictions are also a danger in their own right if they are treated as reliable, for example when published as election forecasts. This is one reason we report absolute performance levels prominently and characterise our results as methodological rather than as forecasts. We therefore see the methods studied here as tools for research on inference from non-probability samples, not as substitutes for representative surveys. Moreover, vote choice is protected by the secrecy of the ballot: predictions of individual vote choice must not be used to profile or target identifiable individuals. Our study works exclusively with anonymised, retrospective academic survey data used under the GLES usage agreement,

and we evaluate only aggregate and anonymised individual-level predictive performance.

Acknowledgements

This paper is supported by the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Research, Technology and Space.

References

Georg Ahnert, Max Pellert, David Garcia, and Markus Strohmaier. 2025. Extracting affect aggregates from longitudinal social media data with temporal adapters for large language models. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 15–36. Association for the Advancement of Artificial Intelligence (AAAI).

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3):337–351.

Adam J. Berinsky. 2017. Measuring Public Opinion with Surveys. *Annual Review of Political Science*, 20(1):309–329.

James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. 2024. Synthetic Replacements for Human Survey Data? The Perils of Large Language Models. *Political Analysis*, 32(4):401–416.

Yong Cao, Haijiang Liu, Arnav Arora, Isabelle Augenstein, Paul Röttger, and Daniel Hershcovich. 2025. Specializing Large Language Models to Simulate Survey Response Distributions for Global Populations. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3141–3154, Albuquerque, New Mexico. Association for Computational Linguistics.

Eric Chu, Jacob Andreas, Stephen Ansolabehere, and Deb Roy. 2023. Language Models Trained on Media Diets Can Predict Public Opinion. *Preprint*, arXiv:2303.16779.

Joshua D Clinton, John S Lapinski, and Marc J Trussler. 2022. Reluctant Republicans, Eager Democrats? Partisan Nonresponse and the Accuracy of 2020 Presidential Pre-election Telephone Polls. *Public Opinion Quarterly*, 86(2):247–269.

Mick P. Couper. 2017. New Developments in Survey Data Collection. *Annual Review of Sociology*, 43(1):121–145.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Die Bundeswahlleiterin. 2017. Bundestagswahl 2017: Ergebnisse Thüringen. `https://www.bundeswahlleiterin.de/bundestagswahlen/2017/ergebnisse/bund-99/land-16.html`. Accessed: 2026-07-27.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. 2024. Questioning the Survey Responses of Large Language Models. In *Advances in Neural Information Processing Systems*, volume 37, pages 45850–45878. Neural Information Processing Systems Foundation, Inc.

Sarah Feldman and Bernard Mendez. 2022. Who Are The People Who Don't Respond To Polls? FiveThirtyEight. Original URL `https://fivethirtyeight.com/features/nonresponse-bias-ipsos-poll-findings/` is defunct following the shutdown of FiveThirtyEight; cited from the Internet Archive snapshot of 26 October 2022.

GLES. 2020. GLES Cross-Section 2009-2017, Cumulation / GLES Querschnitt 2009-2017, Kumulation. ZA6835, Data file Version 1.0.0. German title: GLES Querschnitt 2009-2017, Kumulation.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 540 others. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Hang Jiang, Doug Beeferman, Brandon Roy, and Deb Roy. 2022. CommunityLM: Probing Partisan Worldviews from Language Models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6818–6826, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Damjan Kalajdzievski. 2023. A Rank Stabilization Scaling Factor for Fine-Tuning with LoRA. *Preprint*, arXiv:2312.03732.

Junsol Kim and Byungkyu Lee. 2023. AI-Augmented Surveys: Leveraging Large Language Models and Surveys for Opinion Prediction. *Preprint*, arXiv:2305.09620.

Markus Klein. 2014. Gesellschaftliche Wertorientierungen, Wertewandel und Wählerverhalten. In Jürgen W. Falter and Harald Schoen, editors, *Handbuch Wahlforschung*, pages 563–590. Springer Fachmedien, Wiesbaden.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, and 31 others. 2023. Holistic Evaluation of

Language Models. *Transactions on Machine Learning Research*.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A. Raffel. 2022. Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning. *Advances in Neural Information Processing Systems*, 35:1950–1965.

Yinan Mei, Shaoxu Song, Chenguang Fang, Haifeng Yang, Jingyun Fang, and Jiang Long. 2021. Capturing Semantics for Imputation with Pre-trained Language Models. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, pages 61–72.

Meta AI. 2024. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. Meta AI Blog.

Kevin J. Mullinix, Thomas J. Leeper, James N. Druckman, and Jeremy Freese. 2015. The Generalizability of Survey Experiments. *Journal of Experimental Political Science*, 2(2):109–138.

Avanika Narayan, Ines Chami, Laurel Orr, and Christopher Ré. 2022. Can Foundation Models Wrangle Your Data? *Proceedings of the VLDB Endowment*, 16(4):738–746.

Yao Qu and Jue Wang. 2024. Performance and biases of Large Language Models in public opinion simulation. *Humanities and Social Sciences Communications*, 11(1):1095.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. Qwen2.5 Technical Report. *Preprint*, arXiv:2412.15115.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose Opinions Do Language Models Reflect? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.

Rüdiger Schmitt-Beck, Sigrid Roßteutscher, Harald Schoen, Bernhard Weßels, and Christof Wolf. 2022. A New Era of Electoral Instability. In Rüdiger Schmitt-Beck, Sigrid Roßteutscher, Harald Schoen, Bernhard Weßels, and Christof Wolf, editors, *The Changing German Voter*, pages 3–24. Oxford University Press.

Harald Schoen, Sigrid Roßteutscher, Rüdiger Schmitt-Beck, Bernhard Weßels, and Christof Wolf, editors. 2017. *Voters and Voting in Context: Multiple Contexts and the Heterogeneous German Electorate*. Oxford University Press, Oxford.

Hidetoshi Shimodaira. 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244.

Eleanor Singer and Richard A. Kulka. 2002. Paying Respondents for Survey Participation. In Michele Ver Ploeg, Robert A. Moffitt, and Constance F. Citro, editors, *Studies of Welfare Populations: Data Collection and Research Issues*, chapter 4, pages 105–128. National Academy Press, Washington, DC.

Joseph Suh, Erfan Jahanparast, Suhong Moon, Minwoo Kang, and Serina Chang. 2025. Language Model Fine-Tuning on Scaled Survey Data for Predicting Distributions of Public Opinions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21147–21170, Vienna, Austria. Association for Computational Linguistics.

Leah von der Heyde, Anna-Carolina Haensch, and Alexander Wenz. 2025. Vox Populi, Vox AI? Using Large Language Models to Estimate German Vote Choice. *Social Science Computer Review*, 44(3):549–571.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024. "My Answer is C": First-Token Probabilities Do Not Match Text Answers in Instruction-Tuned Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7407–7416, Bangkok, Thailand. Association for Computational Linguistics.

Makoto Yamada, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Masashi Sugiyama. 2011. Relative density-ratio estimation for robust distribution comparison. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, NIPS'11, pages 594–602, Red Hook, NY, USA. Curran Associates Inc.

A Importance Weighting of Classifier Baselines

To assess whether the tabular classifier baselines in the convenience sample experiments (E2–E4) can be improved by correcting for covariate shift, we apply importance weighting (Shimodaira, 2000). The density ratio $w(x) = p_{\text{test}}(x) \,/\, p_{\text{train}}(x)$ is estimated directly using RuLSIF (Relative unconstrained Least-Squares Importance Fitting) (Yamada et al., 2011), a kernel-based method that avoids separate density estimation for numerator and denominator. Training samples are then reweighted by the estimated density ratio and passed to the classifier via the `sample_weight` parameter.

		Macro F1		TVD	
		Vanilla	IW	Vanilla	IW
Exp.	Model				
E2a	LogReg	0.237 ± 0.019	0.219 ± 0.018	0.295 ± 0.036	0.282 ± 0.036
	RF	0.169 ± 0.022	0.167 ± 0.026	0.322 ± 0.032	0.328 ± 0.035
	CatBoost	0.186 ± 0.019	0.199 ± 0.038	0.349 ± 0.031	0.407 ± 0.068
E2b	LogReg	0.158 ± 0.015	0.161 ± 0.013	0.306 ± 0.034	0.294 ± 0.036
	RF	0.122 ± 0.027	0.109 ± 0.013	0.337 ± 0.009	0.333 ± 0.019
	CatBoost	0.139 ± 0.015	0.135 ± 0.020	0.362 ± 0.020	0.427 ± 0.047
E3a	LogReg	0.311 ± 0.017	0.304 ± 0.012	0.156 ± 0.026	0.153 ± 0.028
	RF	0.325 ± 0.010	0.330 ± 0.010	0.126 ± 0.013	0.123 ± 0.015
	CatBoost	0.286 ± 0.016	0.327 ± 0.015	0.138 ± 0.030	0.160 ± 0.019
E3b	LogReg	0.212 ± 0.022	0.202 ± 0.025	0.154 ± 0.017	0.144 ± 0.020
	RF	0.199 ± 0.020	0.180 ± 0.016	0.132 ± 0.018	0.138 ± 0.016
	CatBoost	0.167 ± 0.018	0.211 ± 0.013	0.144 ± 0.021	0.167 ± 0.012
E4a	LogReg	0.154 ± 0.027	0.136 ± 0.028	0.315 ± 0.040	0.278 ± 0.047
	RF	0.130 ± 0.033	0.133 ± 0.029	0.299 ± 0.051	0.302 ± 0.047
	CatBoost	0.089 ± 0.011	0.165 ± 0.031	0.325 ± 0.065	0.333 ± 0.062
E4b	LogReg	0.107 ± 0.012	0.109 ± 0.016	0.326 ± 0.041	0.301 ± 0.043
	RF	0.094 ± 0.018	0.094 ± 0.013	0.317 ± 0.039	0.323 ± 0.048
	CatBoost	0.070 ± 0.010	0.105 ± 0.008	0.327 ± 0.066	0.357 ± 0.055

Table 1: Macro-F1 and TVD (mean ± 95% CI) for vanilla and importance-weighted (IW) classifier baselines across convenience sample experiments E2–E4.

Table 1 shows that importance weighting has mixed effects across models. For logistic regression and random forest, macro-F1 and TVD remain largely unchanged, with overlapping confidence intervals. CatBoost, however, shows notable macro-F1 improvements with importance weighting in E3 and E4 (e.g., E4a: $0.089 \rightarrow 0.165$), likely because the reweighting helps its more flexible decision boundaries generalise better. At the same time, CatBoost's TVD increases in E2, indicating that improved per-sample accuracy does not necessarily translate to better aggregate vote share estimation. Overall, even where importance weighting helps, the resulting performance remains well below that of fine-tuned LLMs, suggesting that the gap is not primarily caused by covariate shift in the feature space.

B Robustness Across Model Families

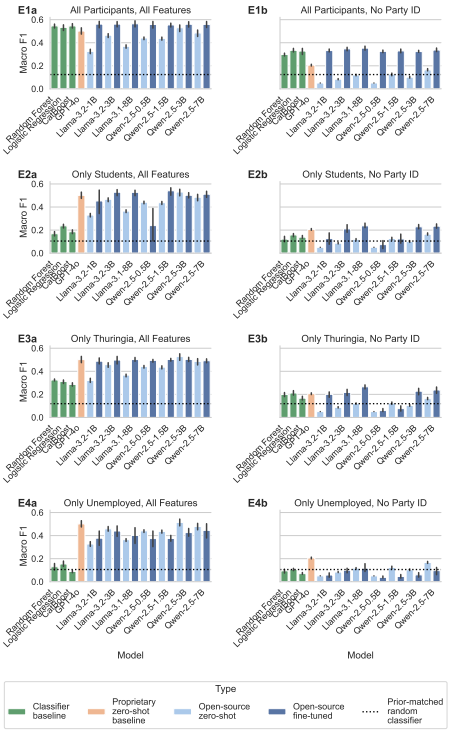

Figure 7: Macro-F1 scores of Llama and Qwen models up to 8B parameters.

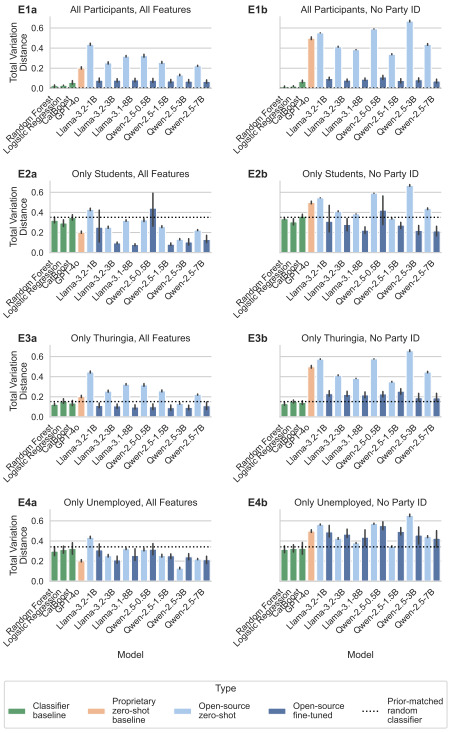

Figure 8: TVD of Llama and Qwen models up to 8B parameters.

C Party Identification Impact

Figure 9 shows permutation feature importance for the logistic regression baseline in E1a: shuffling "party identification" decreases accuracy by roughly 0.31, around four times the drop of the next most important feature (age). The cross-tabulation in Figure 10 illustrates why: most respondents report voting for the party they identify with. This dominant role motivates the "b" variants of our experiments, which remove party identification from the training features.

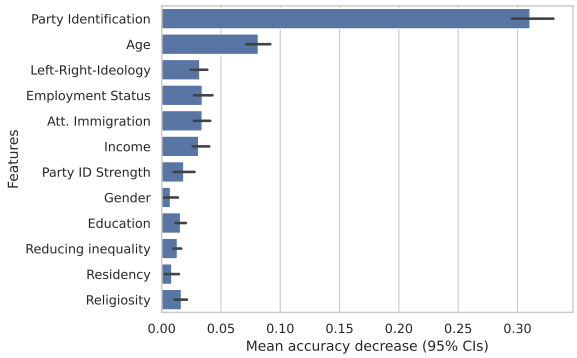

Figure 9: Feature importance for the logistic regression classifier in E1a computed using feature permutation, where each feature's values are randomly shuffled to quantify the drop in performance. This drop indicates how much each feature contributes to the model's predictive power.

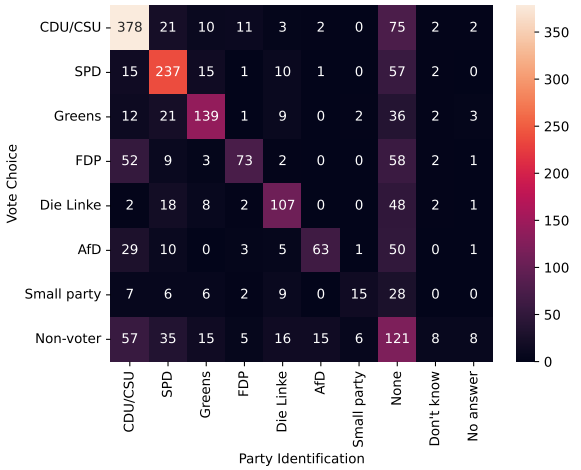

Figure 10: Cross-tabulation of "vote choice" and "party identification".