REASONING BOOSTS OPINION ALIGNMENT IN LLMS

Anonymous authors

Paper under double-blind review

ABSTRACT

Opinion modeling aims to capture individual or group political preferences, enabling applications such as digital democracies, where models could help shape fairer and more popular policies. Given their versatility, strong generalization capabilities, and demonstrated success across diverse text-to-text applications, large language models (LLMs) are natural candidates for this task. However, due to their statistical nature and limited causal understanding, they tend to produce biased opinions when prompted naively. In this work, we study whether reasoning can improve opinion alignment. Motivated by the recent advancement in mathematical reasoning enabled by reinforcement learning (RL), we train models to produce profile-consistent answers through structured reasoning. We evaluate our approach on three datasets covering U.S., European, and Swiss politics. Results indicate that reasoning enhances opinion modeling and is competitive with strong baselines, but does not fully remove bias, highlighting the need for additional mechanisms to build faithful political digital twins using LLMs. By releasing both our method and datasets, we establish a solid baseline to support future research on LLM opinion alignment.

1 Introduction

Could AI give rise to a new kind of democracy, where digital twins vote on our behalf and faithfully reflect our opinions on every issue? Accurate simulations of political behavior offer new opportunities to understand election outcomes and improve policy and democratic processes (Li et al., 2024). However, capturing how diverse individuals reason about political issues remains challenging. While large language models (LLMs) can generate sophisticated political discourse, they often fail to reflect the true diversity of human political viewpoints (Santurkar et al., 2023; Qu & Wang, 2024; Yu et al., 2025; Qi et al., 2024). This raises a fundamental question: how to design agents that reason about politics while faithfully representing human political diversity?

Current approaches to modeling opinions with LLMs predominantly rely on prompting with demographic information and political affiliations, thus making use of learned correlations between demographics and opinions. While these prompt-based methods have grown increasingly sophisticated, they consistently fail to capture real opinion distributions and exhibit unstable, inconsistent responses across different prompts and demographic groups (Santurkar et al., 2023; Ball et al., 2025). We instead seek a preference-consistent approach that directly uses an individual's known opinions rather than demographic proxies.

Yet, the aforementioned methods do not lend themselves to simulating individual political preferences. Methods using richer data sources like interview transcripts can achieve strong performance in modeling individual personas (Park et al., 2024), but remain limited by prohibitive data collection costs, as interviewing is impractical at scale. In this paper, we propose using political survey data as an alternative. Large surveys, like the American National Election Studies (ANES), and Voting Advice Applications (VAAs), i.e., online platforms that match voters with parties or candidates, aim to capture how people position themselves on political issues. While surveys lack the narrative richness and textual signal of interviews, they offer a structured representation of political opinions across populations. Despite their prevalence in bias evaluation (Santurkar et al., 2023; Argyle et al., 2023; Haller et al., 2025) and group-level simulation (Argyle et al., 2023; Cao et al., 2025), surveys have not yet been widely adopted for training individual-level political agents.

Figure 1: **GRPO** for opinion alignment. We use public opinion surveys to align LLMs with individual preference profiles. First, the LLM is fine-tuned with (synthetic) statements and ground-truth answers to adhere to the reasoning template. After fine-tuning, the model answers in the correct format, but is not fully aligned with opinions. We use GRPO with a reward model that rewards proper formatting and correct answers to further improve reasoning.

RL has recently driven major gains in reasoning, with models like o3/o4-mini (OpenAI, 2025) and DeepSeek-R1 (Guo et al., 2025) setting new marks on various benchmarks. Motivated by these results, we use RL, in particular GRPO (Shao et al., 2024), to train agents on survey data, treating opinion formation as a reasoning problem: the agent writes a short rationale and a final choice, and is rewarded when the choice matches the respondent's answer, teaching it to reason toward survey-consistent positions on unseen questions.

We summarize our contributions as follows:

- Reinforcement learning for reasoning-based opinion alignment. We introduce a method that encourages explicit reasoning through reinforcement learning (GRPO) to improve opinion alignment from survey responses.
- Benchmarking on real-world political data. We evaluate our approach on three datasets of authentic political opinions: German parties, U.S. voters, and Swiss candidates, and release them as a benchmark to foster future research. Our results show that reasoning consistently improves opinion alignment.
- Analyzing ideological effects. We investigate how political ideology influences performance by inverting the survey answers of individuals with known positions, revealing systematic variations in alignment quality across the political spectrum.

2 RELATED WORK

Modeling political opinions requires LLMs to adopt specific political personas and answer from these particular viewpoints. Prior work has explored *demographic personas* (representing groups with shared traits like occupation or ethnicity), *character personas* (well-known individuals), and *individualized personas* (digital profiles based on personalized data) (Chen et al., 2024; Tseng et al., 2024). Our focus on modeling individuals makes individualized personas most relevant, though the literature has predominantly employed demographic approaches, with individualized methods represented by only a handful of studies.

Modeling political opinions with demographic personas. Political opinion modeling has primarily used demographic personas through various prompting strategies. Simple approaches use basic political affiliation (e.g., "You are a democrat. What is your opinion on...") (Santurkar et al., 2023), while more complex methods add demographic details like age, education, and location (Argyle et al., 2023; Santurkar et al., 2023; Hwang et al., 2023; Sun et al., 2024; Wang et al., 2025). Recent work has also explored richer contexts, including prompts with demonstrations from belief networks, where models receive related political positions to guide their answer (Chuang et al., 2024). Gudiño et al. (2024) combine demographics with demonstrations, while Stammbach et al. (2024) integrate prompting with odds ratio preference optimization (ORPO) (Hong et al., 2024) to better align outputs with target positions. Despite this methodological diversity, demographic approaches face several limitations.

Table 1: Answers obtained from Llama 3.1 8B trained on an individual in the Right group in ANES. Before training, Llama disagreed; after training, it agreed. Interestingly, the reasoning trace still appeals to similar arguments ("equal opportunity") but frames them differently. Additional examples are provided in Appendix B.

Question:

Do you agree that it is not really that big a problem if some people have more of a chance in life than others?

Before (prompting on base model)

No. I don't think that's right. Everyone deserves an equal shot at succeeding in life, no matter where they're from or who their parents are. If some people have more opportunities and resources, it's not fair to others. We should be working to level the playing field and give everyone a fair chance to succeed.

After (SET+GRPO)

<reasoning> Some inequality in wealth and opportunity is inevitable and indeed necessary to provide
incentives for hard work and innovation – the issue is not eliminating all differences but rather ensuring
that everyone has access to basic necessities and opportunities to improve themselves – other countries with
significant equality have still struggled with poverty and stagnation.

<answer>A) Yes</answer>

Limitations of demographic approaches. Santurkar et al. (2023) identify three critical issues of prompt-based demographic simulation that undermine authentic political representation: *representativeness*, where LLMs' default opinion distributions fail to reflect true population distributions; *steerability*, where demographic prompting provides insufficient control over model outputs; and *consistency*, where political biases vary unpredictably across different topics and contexts. These limitations manifest as systematic biases, with accuracy varying unpredictably across demographic groups (Argyle et al., 2023; Santurkar et al., 2023). While models can approximate overall opinion distributions, they fail to capture natural variance within demographic subgroups and remain

Moving beyond demographics. In contrast to the extensive literature on demographic prompting, few studies have explored representing individual opinion profiles. Park et al. (2024) use interview transcripts to achieve strong accuracy on public opinion surveys, outperforming demographic baselines, but this approach is limited by transcript availability. Haller et al. (2025) propose an in-context

sensitive to prompt variations (Sun et al., 2024; Ball et al., 2025).

learning method, where the model is prompted with respondents' question-answer pairs and then asked to predict the respondents' answers to a set of test questions. However, their focus was bias identification rather than opinion modeling.

Training LLMs to reason. Encouraging LLMs to produce reasoning steps before answering significantly increases accuracy, with approaches like Chain-of-Thought prompting (Wei et al., 2022) or simple prompts like "Let's think step by step" (Kojima et al., 2022). Recent advances have shifted from prompting-based reasoning to training models to generate high-quality reasoning traces through search algorithms or reinforcement learning (Xu et al., 2025). One particular method is Group-relative policy optimization (GRPO) (Shao et al., 2024). GRPO is a variant of Proximal policy optimization (PPO) (Schulman et al., 2017), that uses a group-relative reward normalization rather than learning a value function to estimate advantages. For each prompt, GRPO samples a group of outputs and normalizes rewards by subtracting the group mean and dividing by the group standard deviation. GRPO has proven particularly effective for mathematical reasoning tasks, achieving state-of-the-art results on benchmarks (Guo et al., 2025).

Reasoning for opinion alignment. Yu et al. (2025) propose a persona-based chain-of-thought-inspired prompting framework, which combines demographic, ideological, and temporal factors calibrated using ANES. The authors use this framework to simulate the voting behavior of individual respondents and sum these responses to make aggregate predictions. Unlike our proposed method, they do not use reinforcement learning. Moreover, we focus on learning general preference profiles, while Yu et al. (2025) target voting predictions. While our method could theoretically be applied to the same problem, the computational cost of modeling a representative sample size is currently prohibitive.

Table 2: Dataset overview. Train/test counts refer to questions per unit (candidate/party/respondent).

Setting	Unit	#Units	Labels	Train Q	Test Q
smartvote (CH)	Candidate	18	{Yes, No}	48	12 (topic-strat.)
WoM (DE)	Party	6	{Yes, Neutral, No}	see Table 5	30 (EU&I)
ANES 2020 (US)	Respondent	21	Yes, Neutral, No	67	12 (random)

3 METHOD

We study individual-level opinion alignment from survey data. Consider policy questions such as "Should the government increase funding for renewable energy?" with answer stances like $Y = \{\text{No}, \text{Neutral}, \text{Yes}\}$. Given any policy question q and persona p (a respondent or political party), our goal is to learn the stance of p on q.

3.1 POLITICAL REASONING WITH GRPO

We employ GRPO (Shao et al., 2024) to train agents that generate reasoning traces before providing answers in the following fixed schema:

```
<reasoning>[justification text]</reasoning><answer>[stance]</answer>. (1)
```

Our training data consists of survey responses, $\mathcal{D} = \{(p_j, q_i, y_{ij}^*)\}$, where y_{ij}^* is the observed stance of persona p_j on question q_i . The key challenge is that \mathcal{D} contains no reasoning traces, only final stances. Therefore, our model must learn to generate reasoning that both adheres to the required format above and produces answers that match the ground-truth stances. To achieve this, we use a composite reward function to evaluate and reward each generation along multiple dimensions:

Format Reward (R_{format}): We enforce a structured output format by rewarding correctly placed tags in the schema described in (1), which yields a score of 1 for each of the four tags (maximum of 4).

Length Reward (R_{length}): Let L be the token length of the reasoning trace x_i and L^* desired length. We define a symmetric penalty: $R_{\text{length}} = -|L - L^*|$, which yields a maximum reward of 0 only when the trace length equals L^* .

Correctness Reward (R_{correct}): The primary reward signal comes from matching the survey response: $R_{\text{correct}} = \mathbbm{1}[y_i = y_i^*]$, where y_i is the model's predicted answer and y_i^* is the ground-truth survey response. The total reward is computed as: $R(r_i, y_i^*) = \alpha_1 R_{\text{format}} + \alpha_2 R_{\text{length}} + \alpha_3 R_{\text{correct}}$, with $\alpha_1 = 0.25, \alpha_2 = 0.01, \alpha_3 = 1$.

We do not use an explicit persona representation, and include only a country label in the system prompt. The model is aligned to the individuals solely through correctly answering questions.

Optional Initialization via Supervised Fine-tuning To accelerate convergence and improve initial reasoning quality, we optionally initialize the model through supervised fine-tuning (SFT) before GRPO training. Following (Shao et al., 2024), we construct a dataset of chain-of-thought demonstrations, where each example follows the exact format (1) incentivized by our reward design. The SFT stage serves as a warm start that: (i) reduces the burden on $R_{\rm format}$ during GRPO by pre-training the model to use the correct output structure, and (ii) provides a reasonable initialization for political reasoning.

4 EXPERIMENTS

4.1 DATASETS

For our experiments, we use three datasets from different countries and political systems. The dataset creation process is described in the following paragraphs. We release all data for reproducibility and future research. An overview over the dataset is provided in Table 2.

Table 3: **Reasoning requires large datasets to be effective.** Mean macro-F1 (%) with standard deviations over 8 stochastic runs at T=1.0. For each run, we compute per-unit macro-F1 and then average across units. We report accuracy scores in Appendix E.1. Scores on *smartvote* are considerably better than on *ANES* and *WoM*. This is likely due to the binary nature of *smartvote* and the absence of a <code>Neutral</code> option (see Section 4.6). Performance on *WoM* is better than on *ANES*, likely because of the substantially bigger number of training samples (see also Appendix E.4).

Base model	Method	Dataset			
		smartvote	WoM	ANES	
Untrained baslines	random majority	50.00 37.43	33.33 27.44	33.33 22.98	
Qwen3 1.7B	SFT GRPO SFT+GRPO	55.86 ± 3.04 38.20 ± 7.93 57.30 ± 1.37	41.02 ± 2.75 32.71 ± 3.27 35.54 ± 2.17	35.45 ± 2.35 28.80 ± 1.15 $\mathbf{37.29 \pm 1.11}$	
Llama 3.1 8B	icl SFT GRPO SFT+GRPO	$\begin{array}{c} 55.97 \pm & 5.66 \\ 63.44 \pm & 1.34 \\ 55.14 \pm & 1.25 \\ \textbf{66.88} \pm \textbf{2.18} \end{array}$	$\begin{array}{c} 28.17 \pm 0.02 \\ 48.95 \pm 3.56 \\ 37.29 \pm 2.06 \\ \textbf{52.53} \pm \textbf{4.05} \end{array}$	$\begin{array}{c} 23.20 \pm \ 1.79 \\ \textbf{42.77} \pm \textbf{1.23} \\ 34.55 \pm \ 1.32 \\ 40.66 \pm \ 0.91 \end{array}$	
Qwen3 8B	icl SFT GRPO SFT+GRPO	$\begin{array}{ccc} 60.48 \pm & 4.14 \\ 61.08 \pm & 2.74 \\ 60.64 \pm & 2.01 \\ \textbf{65.11} \pm \textbf{3.33} \end{array}$	$\begin{array}{c} 26.19 \pm 0.35 \\ 42.91 \pm 3.44 \\ 31.42 \pm 1.46 \\ \textbf{49.38} \pm \textbf{1.93} \end{array}$	$\begin{array}{c} 23.20 \pm \ 1.79 \\ 35.14 \pm \ 1.70 \\ 31.47 \pm \ 1.06 \\ \textbf{38.44} \pm \textbf{0.40} \end{array}$	
Magistral 24B	icl SFT GRPO SFT+GRPO	$\begin{array}{c} 66.16 \pm & 0.42 \\ 67.63 \pm & 1.91 \\ 60.56 \pm & 1.93 \\ \textbf{70.73} \pm \textbf{2.21} \end{array}$	$\begin{array}{c} 26.19 \pm 0.35 \\ 51.86 \pm 2.58 \\ 51.00 \pm 3.10 \\ \textbf{53.21} \pm \textbf{3.19} \end{array}$	$\begin{array}{c} 19.23 \pm & 2.91 \\ 39.15 \pm & 0.66 \\ 43.79 \pm & 1.06 \\ \textbf{45.43} \pm \textbf{1.11} \end{array}$	

German party positions For training, we use the official *Wahl-O-Mat* (WoM) dataset (version 26 March 2025) aggregating party positions across federal, state, and European elections (2021–2025). We focus on six major parties (CDU/CSU, SPD, Grüne, FDP, Die Linke, AfD). Each item records agree/neutral/disagree and an explanatory comment. We map to $\{Yes, Neutral, No\}$. For testing, we use EU&I (2024) (Chalkidis, 2024) as an out-of-domain questionnaire (30 items), recoded to the same three labels. Per-party training counts appear in Appendix A.

American National Election Studies From ANES 2020 Time Series (ANES, 2021), we extract 79 policy items (excluding demographics, knowledge, candidate evaluations, and thermometers). Because response formats are heterogeneous, we recode each item to {Yes, Neutral, No} using two schemes detailed in Appendix D: a *conservative* mapping that treats only clearly positive/negative options as agreement/disagreement, and an *aggressive* mapping that collapses most positive responses to Yes. Unless otherwise stated, results use the conservative scheme. Robustness to recoding is analyzed in Appendix E.2. We split items at random into 67 train and 12 test questions. To analyze ideological variation, we sample respondents based on how they place themselves on *ANES*' 7-point liberal—conservative self-placement question. We included 3 respondents per answer option, for a total of 21 respondents.

Swiss candidates We use the English questionnaire for the 2023 Swiss national elections (75 items, of which 60 are policy questions across 12 topics). Responses are on a four-point Likert scale and are collapsed to $\{Yes, No\}$ by mapping $Yes/Rather\ yes \rightarrow Yes$ and $Rather\ no/No \rightarrow No$. We sample three candidates from each of the six parties with at least two National Council seats (18 candidates total). See Appendix A for additional information on these parties. We perform a topic-stratified split: one question per topic held out for test (12), the remaining 48 for training.

Political Spectrum Groups For clarity and comparability across political contexts, we assign parties (*smartvote*, *Wahl-o-Mat*) and ideologies (*ANES*), to three broad political groups: Left, Center, Right. The assignments are detailed in Appendix A.

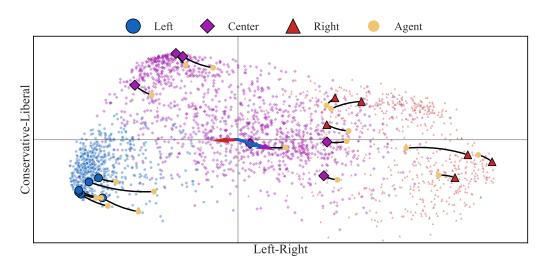


Figure 2: **Agents are more centrist and conservative**. First two dimensions of the principal component analysis of all candidates (small dots) standing for election in the 2023 Swiss national elections. The x- and y-axes correspond to the left-right and conservative-liberal spectra, respectively. The shifts between the positions of the candidates (big dots) included in the *smartvote* dataset and their agents (gold) are depicted by black lines. Unlike results in the literature which indicate a left-libertarian bias (Exler et al., 2025; Hartmann et al., 2023; Rozado, 2024), we observe that our agents are shifted towards the (center-)right. The average distortions between ground-truth and agent position for each group (big arrows) show an overall trend towards more conservative (negative y) and no clear left-right bias. Implementation details are given in Appendix E.5

4.2 EXPERIMENTAL SETUP

We evaluate our approach on different models and against strong baselines. The design targets two questions: (i) how performance scales with model size, and (ii) whether reasoning-pretrained backbones confer advantages over non-reasoning peers. We train one model per unit (candidate/party/respondent). Further training details are given in Appendix F.

Methods We report results for three configurations: GRPO on the survey questions and answers, as well as SFT+GRPO. For *smartvote* and *ANES*, we generate synthetic arguments supporting both positive and negative stances for each policy question using Llama 3.1 70B (Grattafiori et al., 2024) and use those for SFT.

Model backbones We evaluate across four open-weight backbones and sizes to study (i) the effect of model scale and (ii) the impact of prior reasoning pre-training: **Qwen3 1.7B** (Qwen Team, 2025), **Llama3.1 8B** (Grattafiori et al., 2024), **Qwen3 8B** Qwen Team (2025), and **Magistral 24B** Rastogi et al. (2025). Qwen3 and Magistral are pre-trained to reason, allowing us to compare reasoning-pretrained backbones against a non-reasoning counterpart (Llama3.1) at comparable scales. We use 4-bit quantization for all models across all experiments that involve training.

Baselines We report results for two naive baselines. random selects one of the answers uniformly at random. majority answers all questions with the most frequently chosen answer option in that unit's train set. Additionally, we also compare to an in-context learning baseline icl. We follow the methodology described in (Haller et al., 2025), where the LLM is prompted with question-answer pairs and then asked to answer another unseen question. To do so, we report results for the setting where the model receives all the training questions from the same topic area as the test question (or a random subset when limited by context length).

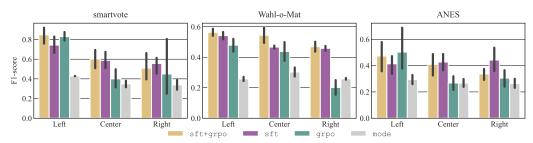


Figure 3: **Not all political positions are equally learnable**. F1 scores reveal that while SFT+GRPO typically works best, every training method underperforms on center and right-leaning groups. Error bars show variance within groups. These disparities may stem from biases baked into Llama 3.1 8B or from inherent differences in how well various political preferences can be learned from survey data. Either way, the results demonstrate that ideology impacts the learnability of political preferences. A more detailed figure with the original parties and ideology groups is presented in Appendix C.

4.3 RESULTS AND ANALYSIS

Table 3 reports macro-F1 averages for all datasets. SFT+GRPO outperforms naive and SFT baselines across datasets, except for Llama 3.1 8B on ANES. Performance varies by dataset: scores peak on smartvote (70.73), while ANES is hardest (45.43). SFT+GRPO performs comparably to or better than icl-baselines on ANES and smartvote, and outperforms them on WoM. Using base models with native reasoning support does not consistently improve performance, as Llama 3.1 8B outperforms Qwen3 8B on WoM and ANES while Magistral 24B performs best overall. GRPO alone underperforms SFT+GRPO, suggesting supervised warm start improves training dynamics. We observe lower performance on datasets with Neutral classes, likely because Neutral aggregates multiple behaviors (uncertainty, social desirability, strategic non-commitment), making it difficult to learn. We further investigate ideology impact in Section 4.5 and Neutral's role in Section 4.6. Results without Neutral appear in Appendix E.1.

4.4 How do agents reason?

In Table 1, we present an example of a reasoning trace. Generally, agents succeed in making coherent arguments for or against a proposed measure that imply the final answer. We identify two failure modes: First, arguments sometimes contain illogical or hallucinated elements. Second, an agent may argue in line with an individual's opinion on one topic, but then fail to do so on other topics. Additional traces and examples of failures are provided in Appendix B.

4.5 POLITICAL IDEOLOGY MATTERS

In Fig. 3, we report F1 by ideology group for Llama 3.1 8B (SFT+GRPO). On *smartvote*, we observe a clear performance disparity between the Left group and the Center and Right groups. *Wahl-o-Mat* and *ANES* show similar patterns, albeit weaker: On *Wahl-o-Mat*, the Left and Center groups perform roughly equally well, while performance is generally lower on *ANES*. For the Right group in *smartvote* and *ANES*, SFT also outperforms all other methods on average, further suggesting difficulties with right-wing preference profiles. These disparities may reflect known left-leaning tendencies in general-purpose LLMs (Hartmann et al., 2023; Exler et al., 2025) or indicate that Right profiles are intrinsically harder to learn from survey signals. The relatively strong Right performance on *WoM* may be aided by larger training sets (cf. Appendix E.4).

4.6 NEUTRAL REASONING POSES A PARTICULAR CHALLENGE

As shown in Fig. 4 (top right), recall is worst on the Neutral class for all individuals in the *ANES* dataset. Furthermore, as can be seen from the regression in Fig. 4 (left column), individuals in the Right group respond with Neutral the most (higher neutral base rate). Clearly, performance for these individuals suffers the most from the difficulties in predicting Neutral. Removing all

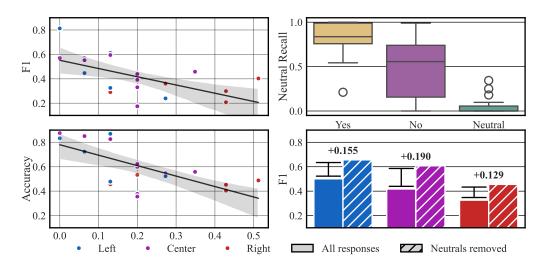


Figure 4: Learning neutral stances is hard. On *ANES*, there is a strong correlation between predictive performance and the neutral base rate on the test set. Individuals in the Right group, and to some extent in the Center group, tend to answer questions with Neutral more often. Both in terms of F1 (top left) and accuracy (bottom left) we observe a negative correlation significant at the 5%-level. Regression details in Appendix E.3. Top right: Recall by class on *ANES*. Averaged over all individuals, the model struggles the most with predicting the Neutral class. Bottom right: Performance in terms of F1 score on *ANES* before (solid) and after (shaded) removing Neutral instances. All groups improve, and the gap between Left and Center becomes narrow. However, the difference between Left and Right remains. While the large number of Neutrals likely depressed performance on the other two classes during training, this could also suggest that the performance disparity is affected by other factors, such as model bias or artifacts from recoding *ANES*. All results were obtained from Llama 3.1 8B trained with SFT+GRPO.

Neutral instances and recomputing the F1 scores leads to a considerable improvement in performance, as shown in Fig. 4 (bottom right), but does not rectify the gap between the Left and Right group. SFT+GRPO also improves relative to the icl-baselines when Neutral is removed (see Appendix E.1). We repeat this analysis under the alternative recoding scheme mentioned in Section 4.1 and further described in Appendix D, and observe the same negative correlation between predictive performance and Neutral base rate, but almost identical performance between the Center and Right groups, indicating that this could, to some extent, be an artifact or recoding ANES. Details are reported in Appendix E.2.

4.7 Whom do agents represent?

To better understand the agents' representativity in semantic terms, we analyze their positions relative to the ground-truth positions of their corresponding human individuals in a two-dimensional model of politics. We perform this on the *smartvote* dataset, for which the interpretation of the principal components is documented (Thurman & Gasser, 2009). Figure 2 presents the positions of the individuals included in the *smartvote* dataset in the space spanned by the first two principal components obtained from the full set of candidates in the 2023 Swiss national elections. Agent positions (gold) were then projected into this space. Not all agents represent their respective candidates (big markers) perfectly. However, despite documented left-libertarian biases of most LLMs (Exler et al., 2025; Hartmann et al., 2023; Rozado, 2024), we do not observe an overall left shift. Instead, the agents' positions are shifted towards the center-right of the left-right (*x*-axis) and the center of the conservative-liberal spectrum (*y*-axis). This is further supported by the group-averaged distances between humans and agents shown in the center of Fig. 2: Right individuals become more left-wing, while Left and Center individuals become more right-wing and conservative.

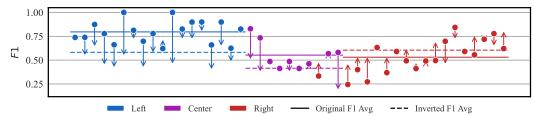


Figure 5: Learning Right and Center positions is harder than Left ones. Points depict F1 scores on the original dataset. Arrows show the difference in F1 scores after all answers in both the train and test sets have been flipped. Left candidates are worse after flipping, both individually and on average. Right candidates improve after flipping, but still perform worse on average than Left candidates. As the disparity is not removed by converting Left to Right, this could suggest that Right positions are inherently more difficult to learn.

4.8 Performance on synthetic positions

To assess how individual preferences affect an associated agent's performance, we create a counterfactual by inverting every *smartvote* answer and training Llama-3.1-8B on the resulting dataset. Inversion roughly corresponds to a 180° rotation about the origin in Fig. 2 (see Appendix E.6), converting, for example, right-wing positions into left-wing ones. For each politician, we compute the F1 difference before and after inversion, as shown in Fig. 5. Members of the Right group benefit from inversion, while the Left and Center groups perform worse on average. Yet even after inversion, the Right does not reach the original F1 of the Left, while the Left drops to roughly the Right 's original level. If the effect were purely due to the model's original political bias, flipping labels should have made right-leaning candidates nearly as easy to predict as left-leaning ones. The fact that this does not occur suggests that left-leaning preference profiles may be intrinsically easier to model, a possibility that merits further investigation.

5 Discussion

Limitations The learning of human opinions is inherently limited by noise, as even the same respondent might answer a question differently on different days. Hence, consistently achieving an F1-score close to 1.0 is very unlikely. The analyses in this paper cover only a small number of individuals or parties, mainly because the method is computationally intensive (i.e., training one model per profile). Data are another constraint: suitable surveys are scarce. Those that exist often have too few items to train individualized agents, or focus on generic themes rather than concrete issues and priorities. Question sets also rarely align across countries, which limits comparability.

Future work The observed disparities between political groups warrant further investigation into bias mitigation. Further research should aim to replace per-agent training with a single model that conditions on compact persona representations (cf. e.g., (Ning et al., 2024)), reducing computational cost and improving data utilization. Real-world deployments will require confidence estimates to gauge answer reliability, achievable by extending the reasoning scheme or through a dedicated value head. Purpose-built questionnaires and improved uncertainty handling may also mitigate the observed difficulties with the Neutral class.

6 CONCLUSION

In this work, we explore individual opinion alignment by training persona-conditioned agents directly on known opinions and by framing answer generation as a structured reasoning task. Across Swiss candidates (*smartvote*), German parties (*Wahl-o-Mat*), and U.S. voters (ANES), SFT+GRPO outperforms prompt-only, in-context, and SFT-only baselines. Performance remains uneven as Neutral stances remain challenging to predict, and accuracy varies significantly across the political spectrum. Our work lays the groundwork and introduces the first benchmark for systematic opinion modeling with LLMs, marking an early step toward the possibility of more representative, AI-driven forms of democracy.

ETHICS STATEMENT

We acknowledge that the modeling of individual opinions in a time of rising mis- and disinformation is not exclusively beneficial, and may pose downstream risks. Precisely because of these risks and the technology's inevitable development, we believe responsible research is essential to investigate how individual opinion alignment can be improved. We use only anonymized survey data and public figures' positions to address privacy concerns. Our research reveals systematic performance disparities across ideological groups, with consistent underperformance on center and right-wing profiles, as well as difficulties modeling neutral positions. We commit to transparent reporting of these biases, as they represent significant system risks that could disenfranchise certain political viewpoints if deployed prematurely. We strongly caution against real-world applications until these fairness issues are resolved, as biased political modeling systems risk undermining democratic representation. Ultimately, we are convinced that with proper bias mitigation, opinion modeling can ultimately enable societally beneficial applications such as digital democracies, where models could help shape fairer and more popular policies.

REPRODUCIBILITY STATEMENT

For reproducibility, we anonymously release our code and datasets, which are linked in Appendix G. Training details, including hyperparameters, are given in Appendix F. Details for all PCA results are described in Appendix E.5.

REFERENCES

American National Election Studies ANES. ANES 2020 Time Series Study Full Release, February 2021.

- Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3):337-351, July 2023. ISSN 1047-1987, 1476-4989. doi: 10.1017/pan.2023.2. URL https://www.cambridge.org/core/journals/political-analysis/article/out-of-one-many-using-language-models-to-simulate-human-samples/035D7C8A55B237942FB6DBAD7CAA4E49.
- Sarah Ball, Simeon Allmendinger, Frauke Kreuter, and Niklas Kühl. Human Preferences in Large Language Model Latent Space: A Technical Analysis on the Reliability of Synthetic Data in Voting Outcome Prediction, February 2025. URL http://arxiv.org/abs/2502.16280. arXiv:2502.16280 [cs].
- Yong Cao, Haijiang Liu, Arnav Arora, Isabelle Augenstein, Paul Röttger, and Daniel Hershcovich. Specializing Large Language Models to Simulate Survey Response Distributions for Global Populations. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *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)*, pp. 3141–3154, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.162. URL https://aclanthology.org/2025.naacl-long.162/.
- Ilias Chalkidis. Investigating LLMs as Voting Assistants via Contextual Augmentation: A Case Study on the European Parliament Elections 2024, October 2024. URL http://arxiv.org/abs/2407.08495. arXiv:2407.08495 [cs].
- Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. From Persona to Personalization: A Survey on Role-Playing Language Agents, October 2024. URL http://arxiv.org/abs/2404.18231. arXiv:2404.18231 [cs].
- Yun-Shiuan Chuang, Krirk Nirunwiroj, Zach Studdiford, Agam Goyal, Vincent V. Frigo, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T. Rogers. Beyond Demographics: Aligning Role-playing LLM-based Agents Using Human Belief Networks, October 2024. URL http://arxiv.org/abs/2406.17232.arXiv:2406.17232 [cs].
- David Exler, Mark Schutera, Markus Reischl, and Luca Rettenberger. Large Means Left: Political Bias in Large Language Models Increases with Their Number of Parameters, May 2025. URL http://arxiv.org/abs/2505.04393. arXiv:2505.04393 [cs].

- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models, 2024.
 - Jairo F. Gudiño, Umberto Grandi, and César Hidalgo. Large language models (LLMs) as agents for augmented democracy. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 382(2285):20240100, November 2024. doi: 10.1098/rsta.2024.0100. URL https://royalsocietypublishing.org/doi/10.1098/rsta.2024.0100. Publisher: Royal Society.
 - Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025.
 - Patrick Haller, Jannis Vamvas, Rico Sennrich, and Lena A. Jäger. Leveraging In-Context Learning for Political Bias Testing of LLMs, June 2025. URL http://arxiv.org/abs/2506.22232. arXiv:2506.22232 [cs].
 - Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation, January 2023. URL http://arxiv.org/abs/2301.01768. arXiv:2301.01768 [cs].
 - Jiwoo Hong, Noah Lee, and James Thorne. ORPO: Monolithic Preference Optimization without Reference Model, March 2024. URL http://arxiv.org/abs/2403.07691. arXiv:2403.07691 [cs].
 - EunJeong Hwang, Bodhisattwa Prasad Majumder, and Niket Tandon. Aligning Language Models to User Opinions, May 2023. URL http://arxiv.org/abs/2305.14929. arXiv:2305.14929 [cs].
 - Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large Language Models are Zero-Shot Reasoners. *Advances in Neural Information Processing Systems*, 35:22199—22213, December 2022. URL https://papers.nips.cc/paper_files/paper/2022/hash/8bb0d29lacd4acf06ef112099c16f326-Abstract-Conference.html.
 - Lincan Li, Jiaqi Li, Catherine Chen, Fred Gui, Hongjia Yang, Chenxiao Yu, Zhengguang Wang, Jianing Cai, Junlong Aaron Zhou, Bolin Shen, Alex Qian, Weixin Chen, Zhongkai Xue, Lichao Sun, Lifang He, Hanjie Chen, Kaize Ding, Zijian Du, Fangzhou Mu, Jiaxin Pei, Jieyu Zhao, Swabha Swayamdipta, Willie Neiswanger, Hua Wei, Xiyang Hu, Shixiang Zhu, Tianlong Chen, Yingzhou Lu, Yang Shi, Lianhui Qin, Tianfan Fu, Zhengzhong Tu, Yuzhe Yang, Jaemin Yoo, Jiaheng Zhang, Ryan Rossi, Liang Zhan, Liang Zhao, Emilio Ferrara, Yan Liu, Furong Huang, Xiangliang Zhang, Lawrence Rothenberg, Shuiwang Ji, Philip S. Yu, Yue Zhao, and Yushun Dong. Political-LLM: Large Language Models in Political Science, December 2024. URL http://arxiv.org/abs/2412.06864. arXiv:2412.06864 [cs].
 - Lin Ning, Luyang Liu, Jiaxing Wu, Neo Wu, Devora Berlowitz, Sushant Prakash, Bradley Green, Shawn O'Banion, and Jun Xie. User-LLM: Efficient LLM Contextualization with User Embeddings, September 2024. URL http://arxiv.org/abs/2402.13598. arXiv:2402.13598 [cs].
 - OpenAI. Introducing OpenAI o3 and o4-mini, April 2025. URL https://openai.com/index/introducing-o3-and-o4-mini/.
 - Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. Generative Agent Simulations of 1,000 People, November 2024. URL http://arxiv.org/abs/2411.10109. arXiv:2411.10109 [cs].
 - Weihong Qi, Hanjia Lyu, and Jiebo Luo. Representation Bias in Political Sample Simulations with Large Language Models, July 2024. URL http://arxiv.org/abs/2407.11409. arXiv:2407.11409 [cs].
 - Yao Qu and Jue Wang. Performance and biases of Large Language Models in public opinion simulation. *Humanities and Social Sciences Communications*, 11(1):1095, August 2024. ISSN 2662-9992. doi: 10.1057/s41599-024-03609-x. URL https://www.nature.com/articles/s41599-024-03609-x. Publisher: Palgrave.
 - Qwen Team. Qwen, April 2025. URL https://qwenlm.github.io/blog/qwen3/.
 - Abhinav Rastogi, Albert Q Jiang, Andy Lo, Gabrielle Berrada, Guillaume Lample, Jason Rute, Joep Barmentlo, Karmesh Yadav, Kartik Khandelwal, Khyathi Raghavi Chandu, et al. Magistral, 2025.
 - David Rozado. The political preferences of LLMs. *PLOS ONE*, 19(7):e0306621, July 2024. ISSN 1932-6203. doi: 10.1371/journal.pone.0306621. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0306621. Publisher: Public Library of Science.

- Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose Opinions Do Language Models Reflect?, March 2023. URL http://arxiv.org/abs/2303.17548.arXiv:2303.17548 [cs].
 - John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms, August 2017. URL http://arxiv.org/abs/1707.06347. arXiv:1707.06347 [cs].
 - Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models, April 2024. URL http://arxiv.org/abs/2402.03300.arXiv:2402.03300 [cs].
 - Dominik Stammbach, Philine Widmer, Eunjung Cho, Caglar Gulcehre, and Elliott Ash. Aligning Large Language Models with Diverse Political Viewpoints, October 2024. URL http://arxiv.org/abs/2406.14155.arXiv:2406.14155 [cs].
 - Seungjong Sun, Eungu Lee, Dongyan Nan, Xiangying Zhao, Wonbyung Lee, Bernard J. Jansen, and Jang Hyun Kim. Random Silicon Sampling: Simulating Human Sub-Population Opinion Using a Large Language Model Based on Group-Level Demographic Information, February 2024. URL http://arxiv.org/abs/2402.18144.arXiv:2402.18144 [cs].
 - James Thurman and Urs Gasser. Three Case Studies from Switzerland: Smartvote. Technical report, Berkman Klein Center, March 2009.
 - Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. Two Tales of Persona in LLMs: A Survey of Role-Playing and Personalization, June 2024. URL https://arxiv.org/abs/2406.01171v3.
 - Yilei Wang, Jiabao Zhao, Deniz S. Ones, Liang He, and Xin Xu. Evaluating the ability of large language models to emulate personality. *Scientific Reports*, 15(1):519, January 2025. ISSN 2045-2322. doi: 10.1038/s41598-024-84109-5. URL https://www.nature.com/articles/s41598-024-84109-5. Publisher: Nature Publishing Group.
 - Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. Advances in Neural Information Processing Systems, 35:24824—24837, December 2022. URL https://proceedings.neurips.cc/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.
 - Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, Chenyang Shao, Yuwei Yan, Qinglong Yang, Yiwen Song, Sijian Ren, Xinyuan Hu, Yu Li, Jie Feng, Chen Gao, and Yong Li. Towards Large Reasoning Models: A Survey of Reinforced Reasoning with Large Language Models, January 2025. URL http://arxiv.org/abs/2501.09686. arXiv:2501.09686 [cs].
 - Chenxiao Yu, Jinyi Ye, Yuangang Li, Zheng Li, Emilio Ferrara, Xiyang Hu, and Yue Zhao. A Large-Scale Simulation on Large Language Models for Decision-Making in Political Science, April 2025. URL http://arxiv.org/abs/2412.15291. arXiv:2412.15291 [cs].

A DATASET DETAILS

In the following, we provide more information on the datasets used in this work.

A.1 SWISS PARTIES

Table 4 lists the Swiss parties included in the *smartvote* dataset, along with their political orientations and group assignments.

Table 4: The six major Swiss parties and their political positions. All parties' candidates answered the same 60 questions (48 train / 12 test).

Party	Political Orientation	Group
SVP	Right-wing conservative	Right
SP	Center-left social democratic	Left
FDP	Center-right liberal	Right
The Center	Center/Christian democratic	Center
Green Party	Left-wing green	Left
GLP	Center-left green liberal	Center

A.2 WAHL-O-MAT

Table 5 lists the German parties included in the *Wahl-o-Mat* dataset, along with their political orientations, group assignments, as well as the train and test set statistics.

Table 5: Number of questions per party in the German dataset across all elections (2021–2025). The test sets contain 30 questions for all parties. A \rightarrow Yes, B \rightarrow No, C \rightarrow Neutral.

Party	Political Orientation	Group	# Train Q.	Train. distr.	Test distr.
CDU/CSU	Center-right Christian democratic	Right	646	Yes Neutral No 52.32 39.16 8.51	Yes Neutral No 60.0 36.67 3.33
SPD	Center-left social democratic	Center	760	Yes Neutral No 55.92 32.63 11.45	Yes Neutral No 63.33 33.33 3.33
Grüne	Center-left green	Left	722	Yes Neutral No 52.08 37.40 10.53	Yes Neutral No 73.33 20.0 6.67
FDP	Center-right liberal	Center	760	Yes Neutral No 48.68 42.37 8.95	Yes Neutral No 46.67 53.33 0.00
Die Linke	Left-wing socialist	Left	646	Yes Neutral No 52.17 41.80 6.04	Yes Neutral No 56.67 40.0 3.33
AfD	Far-right populist	Right	722	Yes Neutral No 45.43 48.20 6.37	Yes Neutral No 30.0 66.67 3.33

A.3 ANES

Table 6 provides the assignment of ANES ideologies to groups.

Table 6: Assignment of ANES ideology self-placement answers to groups.

Ideology	Group
Extremely liberal	Left
Liberal	Left
Slightly liberal	Center
Moderate	Center
Slightly conservative	Center
Conservative	Right
Extremely conservative	Right

B ADDITIONAL REASONING TRACES

In most cases, the reasoning traces are on-topic, align with the final answer, and are internally consistent. In Table 7, we present an example from a German party in the Left group (SPD). Here, the agent is also making use of information it acquired, most likely during the SFT stage on the training set.

Table 7: Example of successful reasoning. The reasoning trace and the final answer are consistent. The agent is also making use of additional information about the party position, most likely acquired during the SFT-stage on the training set.

Question: Social programs should be maintained even at the cost of higher taxes

<reasoning> We are committed to social justice and therefore to strong social work and high social
spending. For us, social policy always means more than mere benefit payments. We are committed to free
childcare for all children up to the age of ten years and for senior citizens and nursing services that look after
them with respect and dignity. /reasoning>

<answer>A</answer>

Agents sometimes resort to rather unconventional arguments. For instance, in the trace in Table 8, an agent trained on a left-wing politician from Switzerland suggested that non-proliferation could suffer if Switzerland were to build new nuclear power plants. While the argument and the reasoning are consistent, it is rather unlikely that this argument would be made.

Table 8: Example of a reasoning trace where the reasoning is formally correct, but the argument made is unlikely to be made in this context.

Question: Should the construction of new nuclear power plants in Switzerland be allowed again?

<reasoning > Switzerland has historically been committed to a policy of non-proliferation of nuclear
weapons. Allowing new nuclear power plants could potentially compromise this policy, as there is a risk that
nuclear materials or technology could be diverted or transferred to other countries with questionable
intentions, potentially contributing to the development of nuclear weapons.
/reasoning >
<answer > B < Answer >

Similarly, in Table 9, the agent starts off in the right direction for the unit it is trained to represent, but then switches topics to prostitution.

Table 9: Example of a trace where the model goes off-topic. This is most likely prompted by the presence of the word "sex".

Question: The legalization of same-sex marriages is a good thing.

<reasoning> In the future, there should be no further debates on the equality of homo and heterosexual
couples. There must also be further prostitution-free areas to protect the neighborhood from the influence of
street prostitution.

<answer>A</answer>

ADDITIONAL FIGURES

756

758

759 760

761

762

768

769

770 771

772

773

774 775 776

777 778 779

781

782

783 784

785

786

787

788

789

790

791

792

793 794

796

797

798 799

800

801

802

804 805

808

809

Figure 6 provides per-party F1 scores, similar to Fig. 3.



Figure 6: Same data as Fig. 3, broken down by party. The party-level data shows a similar trend, SFT+GRPO almost always outperforms SFT and GRPO. F1 tends to be better on left and center-left parties, with the exception of the German center-right FDP.

ANES RECODING SCHEMES D

Questions that ask for (dis-)agreement on a Likert scale were left as is, and the answer options are mapped such that all agreeing options become "Yes" or "Agree", and all disagreeing options become "No" or "Disagree". Options like "Neither agree or disagree" or "None of the above" are mapped to neutral. All other types of questions were rewritten to allow for this answer style. For instance, consider the following original ANES question:

```
When women demand equality these days,
how often are they actually seeking special favors?
[always, most of the time, about half the time,
some of the time, or never / never, some of the time, about half the
time, most of the time, or always]?
1. Always
```

- 2. Most of the time
- 3. About half the time
- 4. Some of the time
- 5. Never

This question can be rewritten as a statement like:

```
When women demand equality these days,
they are actually seeking special favors.
```

For recoding the answers, we consider two schemes:

- Version 1 (conservative): The original options 1 and 2 become "Yes", option 3 "Neutral", and options 4 and 5 are mapped to "No". Generally, options in the middle of the scale like "moderately important", "a moderate amount", "about half of the time", etc. are mapped to "Neutral" under this scheme.
- Version 2 (aggressive): Options 1-4 are considered a "Yes" and only option 5 a "No". Generally, only clearly "negative" options are mapped to "No" under this scheme. There is no neutral option, unless the original options contain one.

The conservative scheme treats only clearly positive frequency responses as agreement, while the aggressive scheme treats any non-zero frequency as agreement.

D.1 FURTHER EXAMPLES

The question

How often do Members of Congress change their votes on legislation because someone donates money to their campaign?

- 1. Never
- 2. Rarely
- 3. A moderate amount of time
- 4. Very often
- 5. All the time

becomes "Do Members of Congress change their votes on legislation because someone donates money to their campaign?" with options "Yes" (4, 5), "Neutral" (3), "No" (1, 2) in version 1, and "Yes" (2,3,4,5) and "No" (1) in version 2.

The question

Does the increasing number of people of many different races and ethnic groups in the United States make this country a better place to live, a worse place to live, or does it make no difference?

- 1. Better
- 833 2. Worse
 - 3. Makes no difference

becomes "Does the increasing number of people of many different races and ethnic groups in the United States make this country a better place?" with options "Yes" (1), "No" (2), and "Neutral" (3) in both versions of the recoding scheme.

E ADDITIONAL MATERIAL

The following sections provide additional material to contextualize the findings in the main text.

E.1 ADDITIONAL RESULT TABLES

Table 10 presents accuracy scores across models and configurations. Similarly, in Table 11, we present accuracy scores without the Neutral class.

E.2 IMPACT OF RECODING ANES

Since the recoding is not unique and can potentially lead to information loss and semantic shifts, we investigate the impact of different recoding schemes on the political groups present in the ANES dataset. Details on the different schemes are provided in Appendix D. The key difference lies in how moderate responses are treated: the conservative scheme maps them to neutral/disagreement, while the aggressive scheme treats most non-zero responses as agreement. We train Llama 3.1 8B on the two recoding schemes and find that mean F1-scores on the aggressive scheme are higher than on the conservative one (45.33 \pm 1.37 vs. 40.67 ± 0.93 reported in Table 3). The confusion matrices in Fig. 7 indicate that the improvement is mostly driven by a lower number of <code>Neutrals</code> in the aggressive scheme. These results highlight that recoding choices can substantially impact both model performance and the political diversity represented in simulated populations.

E.3 REGRESSION TABLES

Table 12 provides the regression table for the regression shown in 4. Similarly, Table 13 provides the regression table for the corresponding analysis shown in Fig. 8.

Table 10: Mean accuracies (%) with standard deviations, averaged over 8 stochastic decoding runs at temperature $T{=}1.0$. For each run, we first compute the per-unit accuracy (candidate/party/respondent) over that unit's test items and then average across units. The table reports the across-run mean \pm s.d. Baselines: random (uniform; 50.0/33.3 on binary/ternary labels by construction) and majority (deterministic). Bold indicates the best method per base model and dataset.

Base model	Method	Dataset			
		smartvote	WoM	ANES	
Untrained baselines	random majority	50.00 37.43	33.33 27.44	33.33 22.98	
Qwen3 1.7B	SFT GRPO SFT+GRPO	61.01 ± 2.88 44.16 ± 0.87 68.17 ± 1.36	51.06 ± 2.66 48.00 ± 3.26 $\mathbf{53.36 \pm 2.57}$	45.63 ± 2.35 38.07 ± 1.24 $\mathbf{52.48 \pm 0.90}$	
Llama 3.1 8B	icl SFT GRPO SFT+GRPO	63.91 ± 3.79 67.23 ± 1.00 66.44 ± 0.86 70.53 ± 1.51	31.44 ± 0.02 68.25 ± 4.30 60.63 ± 1.50 75.06 ± 3.30	$44.39 \pm 1.43 58.82 \pm 0.93 52.35 \pm 1.35 59.21 \pm 0.62$	
Qwen3 8B	icl SFT GRPO SFT+GRPO	$\begin{array}{ccc} 66.09 \pm & 2.85 \\ 65.18 \pm & 1.91 \\ 67.04 \pm & 1.69 \\ \textbf{71.27} \pm \textbf{2.83} \end{array}$	48.67 ± 0.73 61.74 ± 4.63 53.19 ± 2.33 71.16 ± 1.60	41.32 ± 2.98 47.01 ± 1.79 52.92 ± 0.72 $\mathbf{55.14 \pm 0.41}$	
Magistral 24B	icl SFT GRPO SFT+GRPO	71.41 ± 3.54 70.83 ± 1.58 68.44 ± 1.23 73.92 ± 1.89	$\begin{array}{c} 44.64 \pm 0.67 \\ 72.56 \pm 2.40 \\ 72.05 \pm 2.99 \\ \textbf{75.10} \pm \textbf{2.80} \end{array}$	42.46 ± 5.13 52.80 ± 0.94 62.33 ± 0.51 62.33 ± 0.78	



Figure 7: Confusion matrices for Llama 3.1 8B (SFT+GRPO) trained on *ANES* with a more aggressive recoding scheme. $A \rightarrow Yes$, $B \rightarrow No$, $C \rightarrow Neutral$.

E.4 IMPACT OF THE NUMBER OF TRAINING QUESTIONS

We ablate the number of questions in the *WoM* training set for the SFT+GRPO configuration. The results for Llama 3.1 8B are presented in Fig. 9. Increasing the number of training questions is beneficial for all parties, except the center-right FDP party. The FDP's performance may be explained by a shift in the party's positions due to dwindling voter numbers between 2021 and 2025. The largest increase in accuracy is observed for the far-right AfD, which we attribute to the fact that their positions are furthest from the LLM's position, and SFT thus has an aligning or "corrective" effect which becomes stronger as more data is added.

E.5 PCA DETAILS

To obtain the results shown in Fig. 2 and Fig. 10, we compute the PCA as follows: We use the complete *smartvote* dataset for the 2023 national elections across all major parties and all candidates. We convert the original 4-point Likert scale answers to yes-no, as we did for the *smartvote* subset studied in the remainder of the paper. Yes answers are subsequently mapped to 1, and No answers to 0. We then compute the PCA of the resulting matrix, where the rows correspond to candidates, and the columns correspond to questions.

Table 11: Mean accuracies (%) with standard deviations when Neutral is removed. Results are the averages over 8 stochastic decoding runs at temperature $T{=}1.0$. For each run, we first compute the per-unit accuracy (candidate/party/respondent) over that unit's test items and then average across units. The table reports the across-run mean \pm s.d. Baselines: random (uniform; 50.0/33.3 on binary/ternary labels by construction) and majority (deterministic). Bold indicates the best method per base model and dataset (higher is better). smartvote results are unchanged from Table 10 because there are no neutral answers on smartvote.

Base model	Method	Dataset			
		smartvote	WoM	ANES	
Untrained baselines	random majority	50.00 37.43	33.33 27.44	33.33 22.98	
Qwen3 1.7B	SFT GRPO SFT+GRPO	61.01 ± 2.88 44.16 ± 0.87 68.17 ± 1.36	52.66 ± 2.82 49.34 ± 3.11 55.31 ± 2.66	47.76 ± 1.62 40.10 ± 1.21 63.70 ± 1.04	
Llama 3.1 8B	icl SFT GRPO SFT+GRPO	63.91 ± 3.79 67.23 ± 0.10 66.44 ± 0.86 70.53 ± 1.51	31.03 ± 0.19 70.11 ± 4.36 62.93 ± 1.54 78.19 ± 3.38	47.15 ± 1.69 62.47 ± 0.92 64.15 ± 1.51 71.35 ± 0.80	
Qwen3 8B	icl SFT GRPO SFT+GRPO	$\begin{array}{ccc} 66.09 \pm & 2.85 \\ 65.18 \pm & 1.91 \\ 67.04 \pm & 1.69 \\ \textbf{71.27} \pm \textbf{2.83} \end{array}$	31.03 ± 0.19 63.62 ± 4.74 55.30 ± 2.44 73.69 ± 1.60	$\begin{array}{c} 19.11 \pm & 2.93 \\ 54.38 \pm & 2.03 \\ 65.38 \pm & 0.91 \\ \textbf{67.55} \pm \textbf{0.72} \end{array}$	
Magistral 24B	icl SFT GRPO SFT+GRPO	66.09 ± 2.85 70.83 ± 1.58 68.44 ± 1.23 73.92 ± 1.89	46.62 ± 0.80 74.71 ± 2.71 77.21 ± 2.72 74.71 ± 2.71	46.27 ± 5.94 62.67 ± 1.17 66.34 ± 0.55 $\mathbf{69.99 \pm 0.85}$	

Table 12: Regression table for the analysis presented in Fig. 4 and Section 4.6: Model performance vs. Neutral base rate.

Metric	N	Intercept (α)	Slope (β)	Slope SE	Slope 95% CI
F1 Score (Macro)	21	0.5517	-0.6741	0.2091	[-1.3039, -0.4084]
Accuracy	21	0.7803	-0.8562	0.2139	[-1.1117, -0.2364]
Metric	r	R^2	p-value	RMSE	
F1 Score (Macro)	-0.5946	0.3535	0.004472	0.1250	
Accuracy	-0.6763	0.4574	< 0.001	0.1279	

For agents, we replace the ground-truth test set answers with the agents' predictions, and project the resulting vectors onto the first two principal components. Similarly, for the inverted positions shown in Fig. 10, we invert all ground-truth answers, and project the resulting vectors.

E.6 INVERSION

Figure 10 shows candidate positions before and after flipping all answer options. As can be seen, this operation roughly corresponds to a rotation by $180\circ$ about the origin.

F TRAINING DETAILS

In the following, we provide information on training settings and hyperparameters.

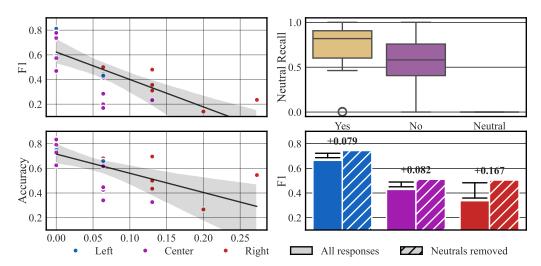


Figure 8: F1 score (A) and accuracy (B) vs. Neutral response rate. Predictive performance decreases significantly as Neutral response rate increases. Comparing the results presented here to those in Fig. 4, we observe the same trend regardless of the coding scheme. (D): Performance in terms of F1 score on ANES before (left) and after (right) removing Neutral instances. All groups improve, but Center only marginally. The gap between Center and Right closes. Performance in predicting Neutral decreased (C). Results were obtained for Llama 3.1 8B using the SFT+GRPO configuration.

Table 13: Regression table for the analysis presented in Fig. 8: Model performance vs. Neutral base rate.

Metric	N	Intercept (α)	Slope (β)	Slope SE	Slope 95% CI
F1 Score (Macro) Accuracy	21 21	$0.6225 \\ 0.7157$	-2.2231 -1.5571	$0.5049 \\ 0.4496$	$ \begin{bmatrix} -3.2799, -1.1662 \\ [-2.4981, -0.6161] \end{bmatrix} $
Metric	r	R^2	p-value	RMSE	
F1 Score (Macro) Accuracy	$-0.7106 \\ -0.6221$	$0.5050 \\ 0.3870$	< 0.001 0.002603	0.1608 0.1431	

F.1 LORA SETTINGS

We use the following LoRA settings for all models and methods. No hyperparameter optimization was performed. $r=32, \alpha=32$, target modules: q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj.

F.2 SYSTEM PROMPT

We used the following system prompt

```
1016
1017
      You are a digital twin of a [NATIONALITY] voter.
1018
      You are asked a policy issue or question.
1019
      You must reason and then answer the question as if you were
1020
      the voter.
1021
      You reason and answer in English.
1022
      Your final answer must be one of the answer options
1023
      ("A" for "Yes", "B" for "No", "C" for "Neutral").
      You must respond in the following format:
1024
      <reasoning>
1025
        Your reasoning goes here.
```



Figure 9: Impact of the number of questions in the training set. Shown are accuracies for 10%, 25%, 50%, 75%, and 100% of available training questions per party on the *WoM* dataset. Overall, F1 increases as the number of training questions increases. The largest increase is observed for the far-right AfD, while the FDP demonstrates a counterexample.

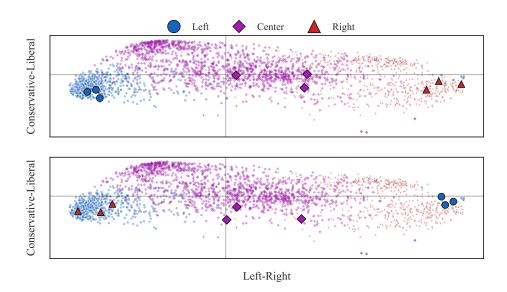


Figure 10: Inverting all answers in the dataset roughly corresponds to a rotation by 180° about the origin. Top: Candidates before inversion. Bottom: Candidates after inversion.

```
</reasoning>
<answer>
Final answer, one of
A) Yes,
B) No,
C) Neutral.
</answer>
```

where we replaced [NATIONALIYT] with "Swiss" for *smartvote*, "German" for *Wahl-o-Mat*, and "US" for *ANES*.

F.3 SFT

We train for 800 steps using the Adam optimizer with a batch size of 8, learning rate 5e-5, a cosine learning rate scheduler, and we fix the maximum gradient norm at 1.0. The first 80 steps were used as warm-up period. For reasoning-pretrained models, we change the <reasoning></reasoning> tags to their respective reasoning templates when fine-tuning. We used the Hugging Face implementation of SFT, and all other hyperparameters were left at the default values.

F.4 GRPO

We train for 800 steps using the Adam optimizer with a batch size of 8, learning rate 5e-6, a cosine learning rate scheduler, and we fix the maximum gradient norm at 1.0. We use the Adam optimizer with $\beta_1=0.9$ and $\beta_2=0.99$, weight decay with 0.1. The first 80 steps were used as a warm-up period. The GRPO group size was 8. The GRPO β parameter was set to 0. The temperature was set at 1.0. For reasoning-pretrained models, we change the reasoning>/reasoning> tags to their respective reasoning templates in the format reward functions. We used the Hugging Face implementation of GRPO, and all other hyperparameters were left at the default values.

F.5 SFT+GRPO

For this configuration, we used the same settings as for the individual SFT and GRPO configurations.

G CODE AND DATA ARTIFACTS

• Code:

https://anonymous.4open.science/r/
Reasoning-Boosts-Opinion-Alignment-in-LLMs-7710/README.md

• Data:

https://huggingface.co/datasets/anonymousemoji/ Reasoning-Boosts-Opinion-Alignment-in-LLMs