Survey Response Generation: Generating Closed-Ended Survey Responses In-Silico with Large Language Models

Georg Ahnert[1], Anna-Carolina Haensch[2,3,4], Barbara Plank[2,3], Markus Strohmaier[1,5,6]
[1]University of Mannheim; [2]LMU Munich; [3]Munich Center for Machine Learning;
[4]University of Maryland, College Park; [5]GESIS Cologne; [6]CSH Vienna
Correspondence: georg.ahnert@uni-mannnheim.de

Abstract

Many *in-silico* simulations of human survey responses with large language models (LLMs) focus on generating closed-ended survey responses, whereas LLMs are typically trained to generate open-ended text instead. Previous research has used a diverse range of methods for generating closed-ended survey responses with LLMs, and a standard practice remains to be identified. In this paper, we systematically investigate the impact that various **Survey Response Generation Methods** have on predicted survey responses. We present the results of 32 mio. simulated survey responses across 8 Survey Response Generation Methods, 4 political attitude surveys, and 10 open-weight language models. We find **significant differences** between the Survey Response Generation Methods in both individual-level and subpopulation-level alignment. Our results show that Restricted Generation Methods perform best overall, and that reasoning output does not consistently improve alignment. Our work underlines the significant impact that Survey Response Generation Methods have on simulated survey responses, and we develop practical recommendations on the application of Survey Response Generation Methods.

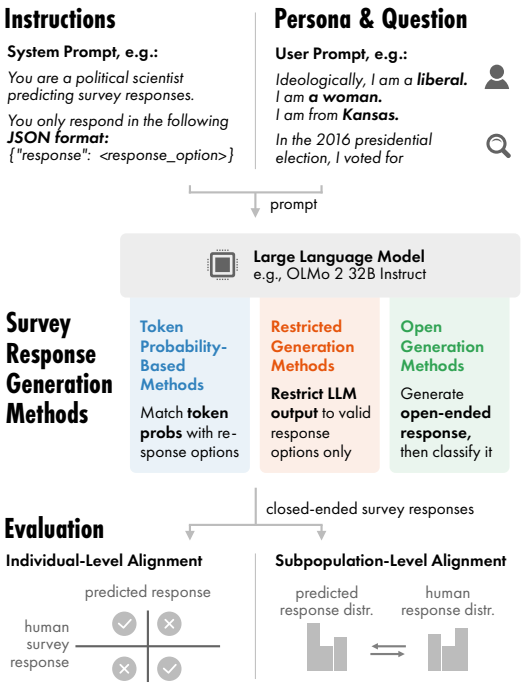


Figure 1: **Survey Response Generation Methods Elicit Closed-Ended Survey Responses From LLMs.** We prompt all models with a combined Persona & Question Prompt to predict political attitudes in the U.S. or Germany. All implemented Survey Response Generation Methods elicit closed-ended survey responses from the LLMs we investigate. We evaluate the individual-level alignment of these responses against human survey data, and the distribution alignment in subpopulations against human response distributions.

1 Introduction

A growing body of research simulates human survey responses by prompting large language models (LLMs) to answer survey questions (Argyle et al., 2023, inter alia). While generative LLMs are designed to generate open-ended text, previous studies have implemented various approaches to constraining LLMs to closed-ended survey responses (Ma et al., 2024). We define **Survey Response Generation Methods** as techniques used to elicit closed-ended responses from large language models to survey questions on attitudes, opinions, and values. Previous research has shown that the closed-ended responses of an LLM can vary strongly from its open-ended responses (Röttger et al., 2024; Wang et al., 2024), but a standard Survey Response Generation Method for social simulations with LLMs has not yet been identified.

In this paper, we evaluate 8 diverse Survey Response Generation Methods against 4 human survey datasets on both individual-level and subpopulation-level alignment, as shown in Figure 1. We focus our evaluation on the **simulation of human survey responses on political attitudes**

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 41554–41577

		Accesses Token-Probabilities	Enforces Format w/ Instructions	Restricts LLM Vocabulary	Generates Open Output First[1]	Generates Probability Distribution
Token Prob.-Based Methods	**First-Token Probabilities**	✔	✔	✗	✗	✔
	First-Token Restricted	✔	✔	✔	✗	✔
	Answer Prefix	✔	✔	✔	✗	✔
Restricted Generation Methods	**Restricted Choice**	✗	✔	✔	✗	✗
	Restricted Reasoning	✗	✔	✔	✔	✗
	Verbalized Distribution	✗	✔	✔	✗	✔
Open Generation Methods	**Open-Ended Classification**	✗	✗[2]	✗[2]	✔	✗
	Open-Ended Distribution	✗	✗[2]	✗[2]	✔	✔

Table 1: **Overview of Survey Response Generation Methods.** Based on previous research (Ma et al., 2024; Röttger et al., 2024, inter alia), we implement a range of Token Probability-Based Methods ■, Restricted Generation Methods ■, and Open Generation Methods ■ for the production of closed-ended survey responses with LLMs. [1]With reasoning models, all methods generate open-ended reasoning output first. [2]The LLM is unrestricted at first, but restricted in its vocabulary and through formatting instructions in the second, classification step.

in the US and Germany and replicate the findings of three influential studies (Argyle et al., 2023; Von Der Heyde et al., 2025; Santurkar et al., 2023) while including novel Survey Response Generation Methods. We find that Restricted Generation Methods perform best, and that reasoning output does not improve performance. Instructing the model to verbalize probabilities for all response options consistently yields the best distributional alignment.

With this study, we contribute to the literature in three ways: (i) we present results from **extensive evaluations** of Survey Response Generation Methods with diverse survey datasets, prompt perturbations, LLMs, and decoding parameters for a total of 32 mio. simulated survey responses. (ii) We highlight the **significant impact of Survey Response Generation Methods** on simulated survey responses, and (iii) we develop **practical recommendations** on which Survey Response Generation Method to use.

2 Survey Response Generation Methods

A growing body of research uses LLMs to simulate human survey responses to questions on attitudes, opinions, and values *in-silico* (Argyle et al., 2023; Park et al., 2024; Boelaert et al., 2025, inter alia). In these studies, an LLM is provided with a description of a persona and a survey question and prompted to predict the survey response of this individual. A large fraction of human survey data is available in a closed-ended format, i.e., each question has a set of response options, categorical or ordinal. Researchers, thus, deploy a diverse range of Survey Response Generation Methods for generating closed-ended survey responses with LLMs.

2.1 Token Probability-Based Methods

This set of Survey Response Generation Methods assumes access to token probabilities and that there are tokens that uniquely identify a single response option. For instance, the token *"Don"* would encode *Donald Trump* as a response option in a question on 2016 U.S. vote choice. We extract its token probability directly from the model output. To increase robustness, the probabilities of tokens that encode the same response option are added up—for instance, *"donald", "Tru"*, etc.

The **First-Token Probabilities Method** ■ is a popular implementation of this approach that extracts probabilities directly on the first output token that a model generates (Argyle et al., 2023; Dominguez-Olmedo et al., 2024; Santurkar et al., 2023; Holtdirk et al., 2025, inter alia).

Since many LLM inference providers only return the top k output tokens and not probabilities over the whole vocabulary, it could happen that some response options do not get a token probability assigned.[1] We therefore additionally implement the **First-Token Restricted Method** ■ that restricts the vocabulary of an LLM to only output tokens from the possible response options.

Wang et al. (2024) showed that the first-token probability of an LLM might not always align with its open-ended response, which instead may start with a prefix, e.g.,*"My answer is "*. One potential mitigation would be to consider token probabilities only after a fixed response-prefix. We implement the prefix above in the **Answer Prefix Method** ■.

[1]e.g., the OpenAI API returns the top 20 tokens only.

2.2 Restricted Generation Methods

This set of methods uses formatting instructions in the system prompt of an LLM to obtain an output that can be easily parsed (Hartmann et al., 2023; Motoki et al., 2023, inter alia). We additionally restrict the vocabulary of the LLM to only the valid response options. While the latter is not strictly necessary for these methods, it ensures that the model follows the formatting instructions provided.

For the **Restricted Choice Method** ■, we dynamically define a JSON schema for each survey question that forces a simple JSON format and only allows for valid response options to be generated by the LLM inside this JSON as follows:

```
{"answer_option": <a valid response option>}
```

The **Restricted Reasoning Method** ■ extends the previous method by first forcing the model to generate reasoning before it generates its choice of response option. The resulting JSON is formatted as follows:

```
{"reasoning": <any string>,
 "answer_option": <a valid response option>}
```

While the previous two methods force the model to generate a single response for each prediction, the **Verbalized Distribution Method** ■ restricts the model to generate a probability distribution over all response options, following Meister et al. (2025). The resulting JSON is formatted as follows:

```
{<response_option_A>: <probability>,
 <response_option_B>: <probability>, ... }
```

Note that we still obtain a distribution over all response options per individual.

2.3 Open Generation Methods

These Survey Response Generation Methods are inspired by a line of work which argues that LLM evaluations with survey questions should be based on open-ended LLM responses (Röttger et al., 2024; Wright et al., 2024; Myrzakhan et al., 2024). Our goal is different, as we aim to simulate human survey responses instead of evaluating the LLM itself (see also Sorensen et al., 2024). Still, for the Open Generation Methods, we do not restrict the output of the model in any way. Instead, we obtain an open-ended response from each LLM and, in a second step, prompt the same model to classify this output according to the survey question and response options at hand. The **Open-Ended Classification Method** ■ implements this two-step approach using the Restricted Choice Method for the classification step, i.e., classifying the output for a single selected response option. The **Open-Ended Distribution Method** ■ uses the Verbalized Distribution Method for the classification step, respectively.

3 Experimental Setup

3.1 Datasets

In this paper, we present a comprehensive evaluation of the above described Survey Response Generation Methods and **compare the obtained predictions to survey responses from human participants.** We focus on political attitudes, as this has been a popular subject for in-silico surveys in the past. Our evaluation spans multiple countries, two languages, and several response option scales, as shown in Table 2. We predict vote choice from the 2016 American National Election Study (ANES, 2016), replicating study 2 from Argyle et al. (2023) with an English-language prompt. We further replicate Von Der Heyde et al. (2025)'s study on vote choice in Germany, predicting responses from the 2017 German Longitudinal Study (GLES, 2017) with a German language prompt. As these two datasets might have leaked into the training data of state-of-the-art LLMs, we also predict self-reported vote choice from the 2025 German federal election (GLES, 2025).[2] Finally, we partially[3] replicate Santurkar et al. (2023) by simulating questions 48–54 from wave 92 of the American Trends Panel (ATP, 2021).

[2] All Llama 3 models and the OLMo training data was published before the survey fieldwork period, and the Qwen models shortly thereafter, so it is unlikely that survey results from the 2025 German election have leaked into the models.

[3] Partial replication of selected questions and with only 1 random seed as we implement multiple, more computationally expensive Survey Response Production Methods.

Survey	#Individuals[1]	#Questions & Topic	Lang. & Country	#Options & Scale Type
ANES 2016	4270	1: vote choice	EN, US	3: categorical
GLES 2017	1976	1: vote choice	DE, DE	9: categorical
GLES 2025	6771	1: vote choice	DE, DE	10: categorical
ATP 2021	500[2]	7: social & cultural change	EN, US	5: ordinal (Likert-scale)

Table 2: **Evaluation Datasets.** We simulate political attitudes across multiple languages, countries, and response scales. [1]Excluding individuals with missing data on the simulated survey questions. [2]Random sample out of 10221 participants in wave 92.

3.2 Prompt Design

To perform predictions for the ANES 2016 and GLES 2017/2025 datasets, we include the same persona attributes and use the same prompt format as Argyle et al. (2023), and Von Der Heyde et al. (2025), only adjusting the year for GLES 2025. For the ATP 2021 dataset, we use Santurkar et al. (2023)'s "bio" format, which is closest to the other prompts. We run all evaluations with the following system prompt: *"You are a political scientist predicting responses to the following question:. . . "*, and include formatting instructions for the Token Probability-Based and Restricted Generation Methods. All prompt templates are provided in Appendix Section B.

3.3 Response Option Scales

To investigate the robustness of all Survey Response Generation Methods against prompt rephrasing, we include 4 variants of the response option scale. The **Full Text** variants consist of the full text of each response option, without using an index—e.g., *'Clinton', 'Trump', 'Non-voter'* for a question on 2016 U.S. vote choice. The **Indexed** variants, sometimes called *multiple choice questions* (Balepur et al., 2025), use an index for the response options instead—e.g., *'A', 'B', 'C'*. Additionally, we evaluate the prompt both in the original order of response options and in a **reversed** order, as previous work has shown option order to be a major source of output variation (Tjuatja et al., 2024; Rupprecht et al., 2025).

3.4 Language Models

We perform our evaluations on **10 open-weight, instruction-tuned and reasoning models** of different sizes from the Llama 3 (3B, 8B, 70B; Llama Team, 2024), OLMo 2 (1B, 7B, 32B; OLMo Team, 2025), and Qwen 3 (8B, 32B; Qwen Team, 2025) families of models—see Appendix Table 8 for the specific model IDs. We also include Qwen 3 8B and Qwen 3 32B with enabled reasoning output in our evaluations. We compare responses obtained from greedy decoding and from the model default temperatures for 3 random seeds.[3] For the Open Generation Methods, we scale temperature for the first, open-ended response step, but keep the default temperature of the model for the classification step. We evaluate the first-token probabilities of reasoning models at the first token after the reasoning output. For computational details, see Appendix A.

3.5 Evaluation

We evaluate all Survey Response Generation Methods by comparing the generated survey responses to human survey data. We include a **stratified baseline** obtained from randomly shuffling the human survey responses in each dataset. As upper bounds of **achievable alignment** on the individual level, we obtain predictions from cross-validations of tuned random forest models. For achievable alignment on the subpopulation level, we use repeated ⅓ sub-sampling (following Suh et al., 2025). We do not expect scores beyond these bounds since human survey responses are not fully predictable.

Individual-Level Alignment First, we calculate the macro avg. F1-score of the generated survey responses against the individual human survey responses. For all methods that generate an individual-level distribution across all response options (see Table 1), we select the most probable response option for evaluation.

Subpopulation-Level Alignment Second, we evaluate the alignment of responses on a subpopulation-level by aggregating individual responses. This is different from previous research that simulated subpopulations directly (e.g., Santurkar et al., 2023). However, simulating individual survey responses first is more versatile as it enables, e.g., individual-level imputation of missing human survey data.

We split the set of respondents into subpopulations by considering all unique values of all persona attributes that were originally included in the simulation by Argyle et al. (2023), Von Der Heyde et al. (2025), and Santurkar et al. (2023), e.g., women & men, people from different states, etc. For *age*, we construct age brackets by flooring to multiples of 10. For all methods that do not generate an individual-level distribution across all response options (see Table 1), we create said distributions through one-hot encoding. We normalize all individual-level distributions to sum up to 1. We then calculate the distribution over response options for a subpopulation as the mean across individual-level distributions. We report the subpopulation-level alignment between the generated distributions and the distribution found in the human survey data for the respective subpopulation using total variation distance for categorical response options (ANES/GLES) and 1-Wasserstein distance for ordinal response options (ATP 2021).

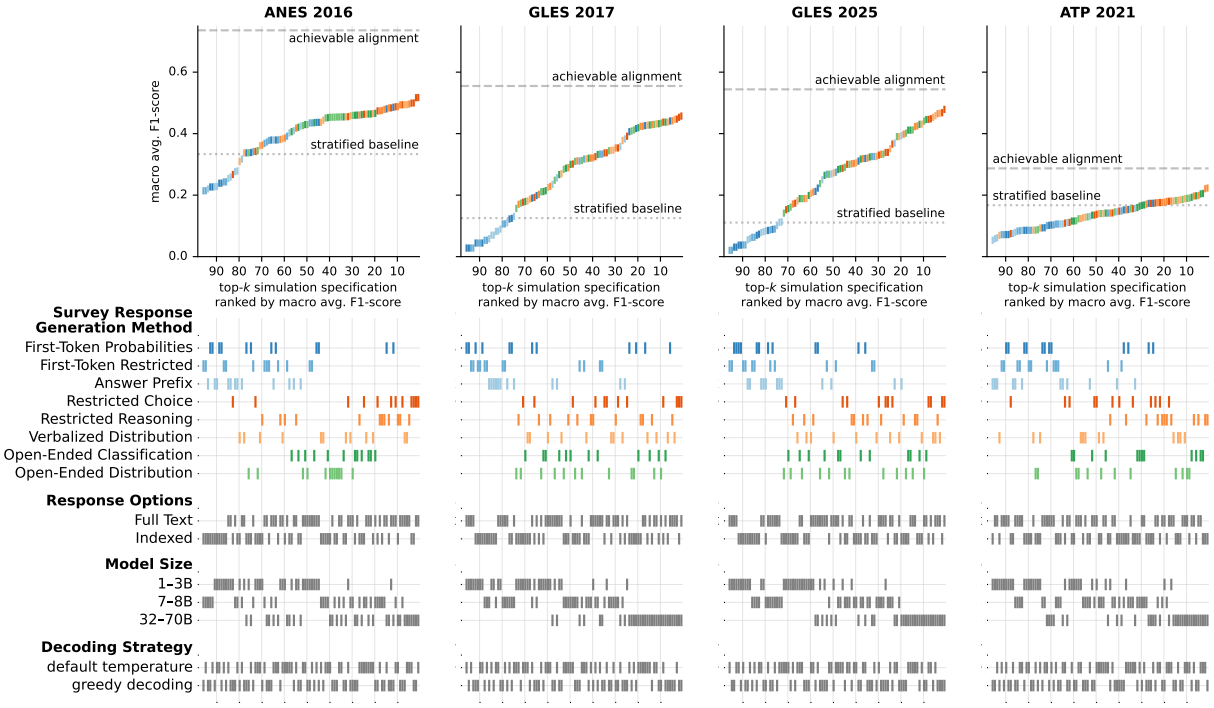


Figure 2: **Individual-Level Alignment Between In-Silico Generated and Human Survey Responses by Dataset (Columns) and Simulation Specification. Top:** macro avg. F1-score (↑) for each aggregated simulation specification, mean across the respective runs. **Bottom:** simulation specification—Survey Response Generation Method, response option variant, model size, and decoding strategy—sorted by macro avg. F1-score (→). Invalid responses are counted as incorrect. **Individual-level alignment varies strongly between Survey Response Generation Methods.** For subpopulation-level alignment, see Appendix Figures 4 & 5.

4 Individual-Level Alignment

First, we view survey response generation as a prediction task and evaluate individual-level alignment. Figure 2 shows macro average F1-scores (top) of aggregated simulation specifications (bottom)—method, response options variant, model size, and decoding strategy.

Survey Response Generation Methods have a large impact on individual-level alignment. Even for specifications that surpass the stratified baseline, we see a difference of > 0.35 on the GLES 2017 and GLES 2025 datasets. Token-Probability Based Methods can yield comparable F1-scores for larger models, but perform poorly for smaller models. This is partially because these methods are more prone to generate invalid responses, especially for reasoning models (see Appendix Figure 10). Restricted Generation Methods, and in particular the Restricted Choice Method, perform best across most datasets and approach the upper bound of achievable alignment on some. LLMs with more parameters also generally outperform smaller models, but we observe no clear pattern for response option scales and decoding strategies.

To investigate the specific impact of Survey Response Generation Methods on individual-level alignment, we fit OLS regressions on each dataset. Table 3 shows the regression coefficients of all Survey Response Generation Methods compared to the First-Token Probabilities Method as a reference, as it is one of the most popular methods. We find that **the Restricted Choice Method leads to significant improvements in individual-level alignment,** followed by the Restricted Reasoning Method and the Open-Ended Classification Method. The Verbalized Distribution Method and the Open-Ended Distribution Method yield significant improvements for some datasets, even if they are designed to generate probability distributions across all response options rather than a single response. We observe similar patterns when using *accuracy* as a metric (see Appendix Table 9). Open Generation Methods generally improve individual-level alignment, but show smaller coefficients for most datasets. This indicates that long, open-ended "reasoning" does not improve the results in our task, while also being orders of magnitude less computationally efficient (see Appendix Figure 3).

	ANES 2016	GLES 2017	GLES 2025	ATP 2021
Intercept	.343*	.037*	.050	.107*
First-Token Restricted	.082*	.051*	.066	-.032*
Answer Prefix	.013	.069*	.059	-.021*
Restricted Choice	**.148***	**.242***	**.218***	.015
Restricted Reasoning	.138*	.219*	.196*	**.052***
Verbalized Distribution	.118*	.233*	.216*	.011
Open-Ended Classif.	.127*	.215*	.184*	.037*
Open-Ended Distrib.	.114*	.212*	.177*	.018*

Table 3: **Regression Coefficients for Individual-Level Alignment (↑).** OLS regression for each dataset with macro avg. F1-score (↑) in each simulation specification as the dependent variable. We use Survey Response Generation Method, response option scale, and LLM as independent variables. We show coefficients for the Survey Response Generation Methods (Reference: First-Token Probabilities ■) and include additional coefficients in Appendix Table 12. Highest coefficient in **bold**, second highest underlined. **The Restricted Choice Method ■ leads to a significant improvements.** * $p < 0.05$, Benjamini-Hochberg adjusted.

	Model Size		
	1–3B	7–8B	32–70B
First-Token Probabilities	0.34	0.23	0.27
First-Token Restricted	0.07	0.15	0.53
Answer Prefix	0.11	0.24	0.32
Restricted Choice	0.18	0.49	**0.74**
Restricted Reasoning	0.15	0.48	**0.69**
Verbalized Distribution	0.07	0.41	**0.60**
Open-Ended Classif.	0.20	0.54	**0.76**
Open-Ended Distrib.	0.14	0.45	**0.69**

Table 4: **Individual-Level Robustness Across Scales.** Mean Fleiss's κ (↑) across all datasets, results with more than 10% invalid values are excluded. We underline $\kappa \geq 0.4$ and use **bold font** for $\kappa \geq 0.6$. **Token Probability-Based Methods ■■■ show poor individual-level robustness across response option scales.** Separate results for each dataset and model, as well as agreement across random seeds are shown in Appendix Figure 7.

4.1 Individual-Level Robustness

In addition to accuracy, we evaluate the robustness of the generated survey responses against prompt perturbations on an individual level. We show the mean agreement between the response option scales—Full Text / Indexed, original / reversed—in Table 4. We observe that especially smaller models (1–3B parameters) often generate disagreeing survey responses. Across all model sizes, Token Probability-Based Methods show little agreement. This indicates that they are subject to biases against certain response options (e.g., A-bias). Out of the Survey Response Generation Methods we evaluate, Restricted Choice and Open-Ended Classification yield the highest agreement across all model sizes, and the Restricted Reasoning Method as well as Open-Ended Distribution are generally not far off.

On one hand, individual-level robustness could be desirable, as the generated survey responses should not be influenced by perturbations of the prompt or of the response option scales. On the other hand, perfect agreement could not be desirable, as persona prompts only provide limited information about the individuals they aim to simulate, and human survey responses are often not uniquely predictable given these attributes. This is exemplified by difficult-to-predict cases.

5 Subpopulation-Level Alignment

While manually investigating the individuals who are the most difficult to predict, we found that their reported vote choice often runs counter to what we would intuitively expect given their other attributes (see Table 5). Evaluating Survey Response Generation Methods with individual-level alignment only does not account for such cases, so we also consider response distributions in subpopulations.

We again fit an OLS regression model using TLV and 1-Wasserstein metrics as the dependent variables. We show the coefficients for the Survey Response Generation Methods in Table 6. We see that **subpopulation-level alignment varies significantly across the Survey Response Generation Methods,** even if simulation specifications

political ideology	party identification	US state	...	true vote choice	predicted vote choice
	a strong Democrat	CA	...	Trump	Clinton
extr. conserv.	a strong Republican	TX	...	Clinton	Trump
conservative	Indep. leaning Rep.	AZ	...	Clinton	Trump
liberal	Indep. leaning Dem.	OH	...	Trump	Clinton
conservative	a strong Republican	NJ	...	Non-Voter	Trump

Table 5: **Most Difficult to Predict Cases in the ANES 2016 dataset,** as identified by a calibrated logistic regression with out-of-fold predictions obtained from 5-fold cross-validation. All five predictions have a true class probability of ≈ 0. See Appendix B for a full list of persona attributes included in the simulation. Given the limited information available, we would not expect an LLM to correctly predict the vote choice reported by these individuals in the ANES survey.

that yield many invalid responses are excluded. We find that the **Verbalized Distribution Method generates well-aligned survey responses** across all datasets. The Restricted Reasoning Method and the Open-Ended Classification Method also yield good subpopulation-level alignment, even if they generate a single survey response and not a distribution over response options for each individual.

Token Probability-Based Methods, on the other hand, often perform worse. This could be explained by Restricted Generation Methods being closer to benchmark-like tasks that LLMs have been trained on during instruction-tuning, while token probabilities are not well aligned for instruction-tuned LLMs—see also Hu et al. (2025) for a similar argument around this *alignment-simulation tradeoff*. Regression coefficients for *Jensen-Shannon divergence* as a dependent variable are presented in Appendix Table 10 and show similar trends.

5.1 Reasoning Models

For both individual-level and subpopulation-level, we find that methods that do not perform open-ended "reasoning"—especially the Verbalized Distribution Method—can yield well-aligned survey responses. This prompts further investigation of two dedicated reasoning models: Qwen 3 8B & Qwen 3 32B. We observe that **reasoning output does not consistently improve subpopulation-level alignment,** and in some cases even degrades alignment—e.g., for the best performing Survey Response Generation Method on the GLES 2025 dataset for Qwen 3 8B: Restricted Reasoning (see Appendix Figure 6b). This is in line with previous findings which show that chain-of-thought reasoning mostly improves model output on mathematical and logic tasks (Sprague et al., 2024). Our evaluation is, however, limited to the default reasoning traces that Qwen 3 8B & Qwen 3 32B are trained to produce and future research could investigate more structured or theory-informed reasoning strategies.

5.2 A Global Perspective On Subpopulation-Level Alignment

Total variation distance and 1-Wasserstein distance compare *in-silico* generated survey responses with human survey responses separately in each subpopulation. The results from all subgroups then have to be aggregated by using a (weighted) average or a regression model, as shown in Table 6. *Distance correlation* is a measure of dependence between random vectors that enables an alternative, global perspective on subpopulation-level alignment (Székely et al., 2007) and has previously been used to compare human to model judgment distributions (Chen et al., 2024). High distance correlation indicates that the dependency structure between subpopulations and survey response distributions in the generated data resembles that in human survey data. We calculate the distance correlation for each simulation specification and show the results in the rightmost column of Table 7 alongside aggregated results for previously discussed individual-level and subpopulation-level assessments. Again, we find that the Restricted Generation Methods yield good global alignment.

	ANES 2016	GLES 2017	GLES 2025	ATP 2021
Intercept	.322*	.476*	.578*	.070*
First-Token Restricted	.105*	.044	-.069	.008
Answer Prefix	.049	-.120*	-.202*	.002
Restricted Choice	.045	-.165*	-.287*	.012*
Restricted Reasoning	.061*	-.197*	-.312*	-.010*
Verbalized Distribution	**-.028**	**-.219***	-.296*	**-.016***
Open-Ended Classif.	.072*	-.174*	-.306*	-.001
Open-Ended Distrib.	.038	-.216*	**-.319***	-.006

Table 6: **Regression Coefficients for Subpopulation-Level Alignment (↓).** OLS regression for each dataset, with total variation distance (↓) as the dependent variable for the ANES and GLES datasets and 1-Wasserstein distance (↓) for the ATP 2021 dataset. Results with more than 10% invalid values were excluded. We use Survey Response Generation Method, response option variant, and LLM as independent variables. We show coefficients for Survey Response Generation Methods (Reference: First-Token Probabilities ■) and include additional coefficients in Appendix Table 13. **Verbalized Distribution ■ leads to significant improvements.** * $p < 0.05$, Benjamini-Hochberg adjusted.

6 Related Work

Generating Survey Responses with LLMs Previous research has investigated how perturbations of the prompt impact the responses of a model to survey questions (Tjuatja et al., 2024; Dominguez-Olmedo et al., 2024; Rupprecht et al., 2025, inter alia). For instance, McIlroy-Young et al. (2024) investigated option-order effects, i.e., the tendency of LLMs to respond to survey questions differently when the order in which the response options are presented is changed. Recently, Cummins (2025) demonstrated that the survey responses generated

	Individual-Level		Subpop.-Level	
	Alignment	Robustness	Alignment	Global Align.
Intercept	-1.647*	-1.183*	-0.930*	-1.150*
First-Token Restrict.	0.751*	0.461*	-0.029	0.613*
Answer Prefix	0.415	0.009	0.667*	-0.073
Restricted Choice	1.535*	1.178*	0.845*	1.138*
Restricted Reasoning	**1.576***	1.097*	1.090*	**1.302***
Verbalized Distrib.	1.443*	0.655*	**1.433***	1.104*
Open-Ended Classif.	1.440*	**1.382***	0.882*	1.281*
Open-Ended Distrib.	1.273*	0.930*	1.097*	1.104*

Table 7: **Regression Coefficients Across All Surveys (normalized, ↑).** OLS regressions for individual-level alignment (macro avg. F1-score), robustness (Fleiss' κ), subpopulation-level alignment (1− total variation distance / 1-Wasserstein distance) and global alignment (distance correlation). Results from each metric are z-score normalized separately per dataset and results with more than 10% invalid values were excluded. We use dataset, Survey Response Generation Method, response option variant, and LLM as independent variables. We show coefficients for Survey Response Generation Methods (Reference: First-Token Probabilities ■) and include additional coefficients in Appendix Table 14. **Restricted Generation Methods ■■■ consistently yield significant improvements** and are computationally efficient. * $p < 0.05$, Benjamini-Hochberg adjusted.

by LLMs can vary widely across input-text specifications and models. In this paper, we instead focus on closed-ended output of LLMs, and still find significant differences between simulation specifications. Future research should combine both prompts and Survey Response Generation Methods for a holistic assessment of how LLMs generate survey responses.

Few papers have investigated the specific impact of Survey Response Generation Methods so far. Wang et al. (2024) identified that the most probable first output token often does not match the open-ended responses of an LLM when prompted with survey questions. Through our assessments, we go beyond dissimilarities in the generated responses and identify the Survey Response Generation Method that works best for a given survey and LLM. Meister et al. (2025) compared different methods for generating response distributions and found that the Verbalized Distribution Method works best, which is why we also implement it. Our work goes further, as we simulate the responses of individual survey participants instead of subpopulations directly. We evaluate Survey Response Generation Methods on non-English datasets (GLES 2017, GLES 2025), and compare a wider range of methods, including Open Generation Methods.

Response Generation in Other Settings Another line of research has investigated closed-ended and open-ended model responses in other contexts. Röttger et al. (2024) found that adding instructions on the response format, and forcing models to, e.g., 'take a clear stance', alters the response option that a model chooses. Tam et al. (2024) also observed a negative impact of format instructions for mathematical reasoning tasks. Finally, Licht et al. (2025) evaluate methods for annotating scalar constructs with LLMs and see improvements with pairwise comparisons and token probability-weighted scores. We extend this line of research to the simulation of survey responses, including persona prompts and human survey data for comparison.

7 Recommendations and Conclusion

In this paper, we present a **systematic assessment of Survey Response Generation Methods** for generating closed-ended survey responses *in-silico* with large language models. Our evaluations span 8 Survey Response Generation Methods, 4 political attitude datasets across 2 countries and languages, and 10 open-weight LLMs, as well as multiple robustness checks.[4]

Recommendations: (i) We argue that the **choice of Survey Response Generation Method should be well-justified** for *in-silico* surveys since we find significant differences between these methods. (ii) We **do not recommend the use of Token Probability-Based Methods ■■■,** as they generate misaligned survey responses. (iii) For predicting closed-ended survey responses, we suggest to **consider Restricted Generation Methods ■■■ first,** as they consistently showed significant improvement over other methods while also being more computationally efficient than Open Generation Methods ■■.

This paper should serve as a starting point for future research on how to generate valid, reliable, and useful survey responses with LLMs.

[4]Code & Data: https://github.com/dess-mannheim/survey_response_generation

Limitations

Our main focus in this paper lies on the evaluation of Survey Response Generation Methods for the *in-silico* simulation of surveys on political attitudes. We include a diverse range of datasets across countries with different political systems and languages and aim to replicate influential studies on *in-silico* surveys Argyle et al. (2023); Von Der Heyde et al. (2025); Santurkar et al. (2023). Many of our overall findings generalize across these datasets. Still, further evaluation should be performed in non-Western contexts and on surveys of attitudes, opinions, and values that go beyond the topics that we have studied. The already large variety of simulation specifications we have included—Survey Response Generation Methods, LLM, response option scale, decoding strategy, etc.—also created computational constraints that did not allow us to investigate, for instance, the impact of language and country/political context independently between the ANES and GLES datasets.

Future research should also investigate further perturbations to the response options scales such as a missing midpoint option, which have been found to impact human and LLM survey responses (Tjuatja et al., 2024; Rupprecht et al., 2025; McIlroy-Young et al., 2024). Other applications of closed-ended questions, e.g., for LLM benchmarking or to investigate model alignment, are beyond the scope of this paper, and it remains to be demonstrated whether our results generalize to these contexts as well.

We define Survey Response Generation Methods as being concerned with how closed-ended survey responses are generated by and can be extracted from LLMs. This specifically limits the scope of our paper to exclude different approaches to prompting LLMs for survey simulation (Lutz et al., 2025; Park et al., 2024). Instead, we adopt a persona prompt template from each study that we replicate (Argyle et al., 2023; Von Der Heyde et al., 2025; Santurkar et al., 2023). For each model and dataset, we evaluate all Survey Response Generation Methods using 4 response option scale variants, {Full Text, Indexed} $\times$ {original order, reversed order}, and 3 random seeds[5] to investigate the robustness of our findings. Still, alternative persona prompts (e.g., interview-style) have been shown to positively influence the LLM response alignment (Lutz et al., 2025) and might interact differently with different Survey Response Generation Methods. Finally, while we do investigate the joint impact of `temperature` and `top-k` for a subset of the Survey Response Generation Methods, more advanced decoding strategies (Zhang et al., 2024; Garces Arias et al., 2025, inter alia) are also out of the scope of this paper.

We note that treating human survey responses as ground truth ignores biases in human survey response (Groves and Lyberg, 2010). Future research should therefore consider evaluations of *in-silico* survey responses that can be performed without this assumption.

[5]For the ATP 2021 dataset, we only used 1 seed but include responses to 7 different questions.

Ethical Considerations

Survey Response Generation Methods are frequently under-reported or insufficiently described in existing research, which poses challenges for reproducibility and transparency. In-silico surveys represent an emerging and active area of inquiry within both NLP and survey methodology. These approaches hold considerable promise for advancing survey research, for example to pre-test survey instruments or to impute missing data.

However, uncritical applications—particularly those aimed at directly predicting survey outcomes—carry the risk of distorting public opinion, with a heightened potential to misrepresent the perspectives of marginalized populations. Furthermore, unresolved epistemological questions persist regarding the extent to which simulated survey responses can meaningfully inform our understanding of the populations they aim to represent. Finally, issues of inferential privacy warrant careful consideration, as individuals may not consent to the simulation of their responses, particularly in cases where they have deliberately chosen not to participate in surveys.

Acknowledgments

We would like to thank Marlene Lutz, Tobias Schumacher, Indira Sen, Florian Lemmerich, Florian Keusch, Frauke Kreuter, and the members of the FK^{2}RG research meeting for their helpful feedback on earlier versions of this project.

References

ANES. 2016. 2016 Time Series Study.

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.

ATP. 2021. The American Trends Panel.

Nishant Balepur, Rachel Rudinger, and Jordan Lee Boyd-Graber. 2025. Which of These Best Describes Multiple Choice Evaluation with LLMs? A) Forced B) Flawed C) Fixable D) All of the Above. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3394–3418, Vienna, Austria. Association for Computational Linguistics.

Julien Boelaert, Samuel Coavoux, Étienne Ollion, Ivaylo Petev, and Patrick Präg. 2025. Machine Bias. How Do Generative Language Models Answer Opinion Polls? *Sociological Methods & Research*, 54(3):1156–1196.

Beiduo Chen, Xinpeng Wang, Siyao Peng, Robert Litschko, Anna Korhonen, and Barbara Plank. 2024. "Seeing the Big through the Small": Can LLMs Approximate Human Judgment Distributions on NLI from a Few Explanations? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14396–14419, Miami, Florida, USA. Association for Computational Linguistics.

Jamie Cummins. 2025. The threat of analytic flexibility in using large language models to simulate human data: A call to attention. *arXiv preprint*. ArXiv:2509.13397 [cs].

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.

Yixin Dong, Charlie F. Ruan, Yaxing Cai, Ruihang Lai, Ziyi Xu, Yilong Zhao, and Tianqi Chen. 2025. XGrammar: Flexible and Efficient Structured Generation Engine for Large Language Models. *arXiv preprint*. ArXiv:2411.15100 [cs].

Esteban Garces Arias, Meimingwei Li, Christian Heumann, and Matthias Assenmacher. 2025. Decoding Decoded: Understanding Hyperparameter Effects in Open-Ended Text Generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9992–10020, Abu Dhabi, UAE. Association for Computational Linguistics.

GLES. 2017. GLES 2017 Post-Election Cross Section.

GLES. 2025. GLES 2025 Post-Election Cross Section.

R. M. Groves and L. Lyberg. 2010. Total Survey Error: Past, Present, and Future. *Public Opinion Quarterly*, 74(5):849–879.

Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. 2023. The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation.

Tobias Holtdirk, Dennis Assenmacher, Arnim Bleier, and Claudia Wagner. 2025. Learning from Convenience Samples: A Case Study on Fine-Tuning LLMs for Survey Non-response in the German Longitudinal Election Study. *arXiv preprint*. ArXiv:2509.25063 [cs].

Tiancheng Hu, Joachim Baumann, Lorenzo Lupo, Nigel Collier, Dirk Hovy, and Paul Röttger. 2025. SimBench: Benchmarking the Ability of Large Language Models to Simulate Human Behaviors. *arXiv preprint*. ArXiv:2510.17516 [cs].

Maximilian Kreutner, Jens Rupprecht, Georg Ahnert, Ahmed Salem, and Markus Strohmaier. 2026. QSTN: A Modular Framework for Robust Questionnaire Inference with Large Language Models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 537–549, Rabat, Marocco. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626, Koblenz Germany. ACM.

Hauke Licht, Rupak Sarkar, Patrick Y. Wu, Pranav Goel, Niklas Stoehr, Elliott Ash, and Alexander Miserlis Hoyle. 2025. Measuring scalar constructs in social science with LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32132–32159, Suzhou, China. Association for Computational Linguistics.

Llama Team. 2024. The Llama 3 Herd of Models. *arXiv preprint*. ArXiv:2407.21783 [cs].

Marlene Lutz, Indira Sen, Georg Ahnert, Elisa Rogers, and Markus Strohmaier. 2025. The Prompt Makes the Person(a): A Systematic Evaluation of Sociodemographic Persona Prompting for Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23212–23237, Suzhou, China. Association for Computational Linguistics.

Bolei Ma, Xinpeng Wang, Tiancheng Hu, Anna-Carolina Haensch, Michael A. Hedderich, Barbara Plank, and Frauke Kreuter. 2024. The Potential and Challenges of Evaluating Attitudes, Opinions, and Values in Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP*

2024, pages 8783–8805, Miami, Florida, USA. Association for Computational Linguistics.

Reid McIlroy-Young, Katrina Brown, Conlan Olson, Linjun Zhang, and Cynthia Dwork. 2024. Order-Independence Without Fine Tuning. *Advances in Neural Information Processing Systems*, 37:72818–72839.

Nicole Meister, Carlos Guestrin, and Tatsunori Hashimoto. 2025. Benchmarking Distributional Alignment of Large Language Models. 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 24–49, Albuquerque, New Mexico. Association for Computational Linguistics.

Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2023. More human than human: measuring ChatGPT political bias. *Public Choice*, 198:3–23.

Aidar Myrzakhan, Sondos Mahmoud Bsharat, and Zhiqiang Shen. 2024. Open-LLM-Leaderboard: From Multi-choice to Open-style Questions for LLMs Evaluation, Benchmark, and Arena. *arXiv preprint*. ArXiv:2406.07545 [cs].

OLMo Team. 2025. 2 OLMo 2 Furious. *arXiv preprint*. ArXiv:2501.00656 [cs].

Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. 2024. Generative Agent Simulations of 1,000 People. *arXiv preprint*.

Qwen Team. 2025. Qwen3 Technical Report. *arXiv preprint*. ArXiv:2505.09388 [cs].

Jens Rupprecht, Georg Ahnert, and Markus Strohmaier. 2025. Prompt Perturbations Reveal Human-Like Biases in LLM Survey Responses. *arXiv preprint*. ArXiv:2507.07188 [cs].

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schuetze, and Dirk Hovy. 2024. Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15295–15311, Bangkok, Thailand. Association for Computational Linguistics.

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*, pages 29971–30004. PMLR.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. Position: a roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pages 46280–46302, Vienna, Austria. JMLR.org.

Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2024. To CoT or not to CoT? Chain-of-thought helps mainly on math and symbolic reasoning. *arXiv preprint*. ArXiv:2409.12183 [cs].

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. *arXiv preprint*. ArXiv:2502.16761 [cs].

Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. 2007. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6).

Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. 2024. Let Me Speak Freely? A Study On The Impact Of Format Restrictions On Large Language Model Performance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1218–1236, Miami, Florida, US. Association for Computational Linguistics.

Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkwar, and Graham Neubig. 2024. Do LLMs Exhibit Human-like Response Biases? A Case Study in Survey Design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026.

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*.

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.

Dustin Wright, Arnav Arora, Nadav Borenstein, Srishti Yadav, Serge Belongie, and Isabelle Augenstein. 2024. LLM Tropes: Revealing Fine-Grained Values and Opinions in Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 17085–17112, Miami, Florida, USA. Association for Computational Linguistics.

Jianyi Zhang, Da-Cheng Juan, Cyrus Rashtchian, Chun-Sung Ferng, Heinrich Jiang, and Yiran Chen. 2024. SLED: Self Logits Evolution Decoding for Improving Factuality in Large Language Models. In *Advances in Neural Information Processing Systems*, volume 37, pages 5188–5209. Curran Associates, Inc.

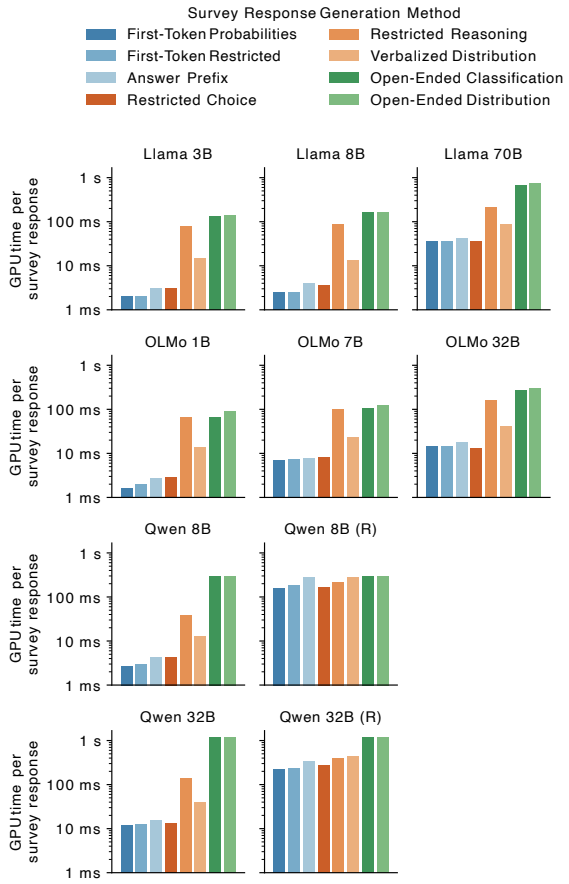


Figure 3: **Mean GPU Time for A Single Survey Response.** We run all models using vllm on 2 NVIDIA H100 GPUs (tensor-parallel). We report GPU time instead of token count to accommodate for optimizations such as automatic prefix caching, but also for the overhead that is created by restricting the vocabulary of an LLM with structured outputs. Considering the log-scale y-axis, Open Generation Methods, larger models, and in particular reasoning models require **orders of magnitude more GPU time** than Token Probability-Based Methods, or the Restricted Choice Method.

Short Name	Huggingface Model ID
Llama 3B	`meta-llama/Llama-3.2-3B-Instruct`
Llama 8B	`meta-llama/Llama-3.1-8B-Instruct`
Llama 70B	`meta-llama/Llama-3.3-70B-Instruct`
OLMo 1B	`allenai/OLMo-2-0425-1B-Instruct`
OLMo 7B	`allenai/OLMo-2-1124-7B-Instruct`
OLMo 32B	`allenai/OLMo-2-0325-32B-Instruct`
Qwen 8B	`Qwen/Qwen3-8B`
Qwen 32B	`Qwen/Qwen3-32B`
Qwen 8B (R)	`Qwen/Qwen3-8B with Reasoning`
Qwen 32B (R)	`Qwen/Qwen3-32B with Reasoning`

Table 8: **Language Models.** We evaluate all Survey Response Generation Methods on 10 open-weight LLMs.

A Computational Details

We run all our experiments the 10 open-weight instruction tuned and reasoning models shown in Table 8. For language model inference, we use `vllm` (Kwon et al., 2023) version `0.10.1.1` and the `xgrammar` (Dong et al., 2025) backend for inference with `structured outputs`. We use the `QSTN` framework (Kreutner et al., 2026) to facilitate prompt perturbations and the configuration of structured outputs. We ran all our experiments on 2 NVIDIA H100 GPUs (tensor-parallel). Our experiments for the ANES 2016 dataset had a total runtime of 88h, for the GLES 2017 dataset of 121h, for the GLES 2025 dataset of 297h, and for the ATP 2021 dataset of 108h. We ran our additional experiments on the impact of decoding hyper-parameters (see Figure 8) on 2 NVIDIA RTX PRO 6000 Blackwell GPUs with a total runtime of 29h. To save computational resources, we only generated open-ended output once for each simulation specification (model, seed, response scale, temperature) and then classified the same output separately for the Open-Ended Classification and for the Open-Ended Distribution Methods. Figure 3 shows the average GPU Time spent to generate a single survey response with a given Survey Response Generation Method and LLM.

B Prompts

Tables 15–19 contain all system and user prompts that we used in our evaluations.

C Additional Results

	ANES 2016	GLES 2017	GLES 2025	ATP 2021
Intercept	.450*	.088*	.305*	.204*
First-Token Restricted	.060*	.138*	-.073	-.029*
Answer Prefix	.009	.142*	-.039	-.022
Restricted Choice	.102*	**.298***	**.116***	.006
Restricted Reasoning	**.110***	.289*	.055	**.030***
Verbalized Distribution	.079*	.288*	.105	.001
Open-Ended Classif.	.100*	.266*	.076	.019
Open-Ended Distrib.	.096*	.263*	.061	.010

Table 9: **Regression Coefficients for Individual-Level Accuracy (↑).** OLS regression for each dataset with accuracy (↑) in each simulation specification as the dependent variable. We use Survey Response Generation Method, response option scale, and LLM as independent variables. We show coefficients for the Survey Response Generation Methods (Reference: First-Token Probabilities ■). For macro avg. F1-score as a dependent variable see Table 12. $*p < 0.05$, Benjamini-Hochberg adjusted.

	ANES 2016	GLES 2017	GLES 2025
Intercept	.146*	.228*	.255*
First-Token Restricted	.068*	.105*	.039
Answer Prefix	.032	-.058	-.087*
Restricted Choice	.038	-.082*	-.121*
Restricted Reasoning	.050*	-.112*	-.148*
Verbalized Distribution	**-.040***	**-.131***	-.150*
Open-Ended Classif.	.074*	-.093*	-.140*
Open-Ended Distrib.	.024	-.129*	**-.155***

Table 10: **Regression Coefficients for Subpopulation-Level Jensen-Shannon Divergence (↓).** OLS regression for each dataset, with Jensen-Shannon divergence (↓) as the dependent variable for the ANES and GLES datasets. Results with more than 10% invalid values were excluded. We use Survey Response Generation Method, response option variant, and LLM as independent variables. We show coefficients for Survey Response Generation Methods (Reference: First-Token Probabilities ■). For total variation distance as a dependent variable, see Table 13. $*p < 0.05$, Benjamini-Hochberg adjusted.

	ANES 2016	GLES 2017	GLES 2025	ATP 2021
Intercept	.759*	.779*	.773*	.651*
First-Token Restricted	**.124***	-.058*	.018	-.028
Answer Prefix	-.023	-.101*	-.055	-.030
Restricted Choice	.111*	.053*	.105*	.001
Restricted Reasoning	.117*	**.065***	.117*	**.059***
Verbalized Distribution	.098*	.046*	.105*	.013
Open-Ended Classif.	.123*	.052*	**.123***	.047*
Open-Ended Distrib.	.103*	.040*	.109*	.037*

Table 11: **Regression Coefficients for Subpopulation-Level Distance Correlation (↑).** OLS regression for each dataset, with Distance Correlation (↑) as the dependent variable. Results with more than 10% invalid values were excluded. We use Survey Response Generation Method, response option variant, and LLM as independent variables. We show coefficients for Survey Response Generation Methods (Reference: First-Token Probabilities ■). For total variation distance as a dependent variable, see Table 13. $*p < 0.05$, Benjamini-Hochberg adjusted.

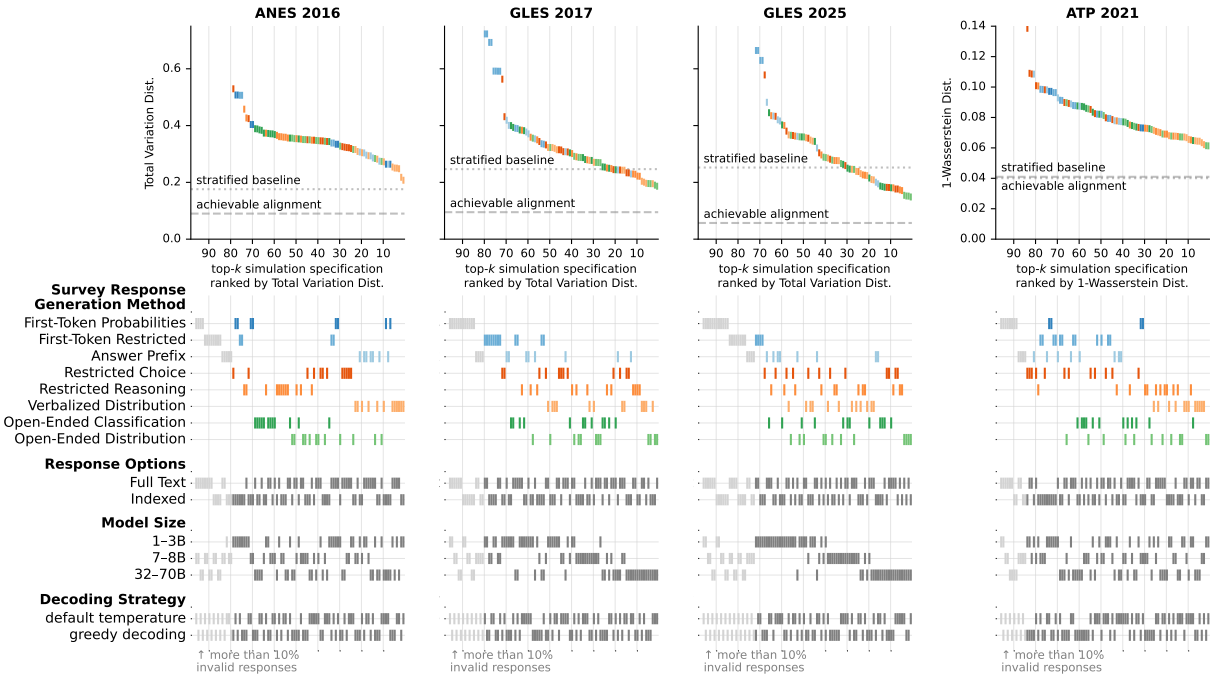


Figure 4: **Subpopulation-Level Alignment: Total Variation Distance/1-Wasserstein Distance.** For the ANES and GLES datasets, we use total variation distance to measure alignment on categorical response options. For the ATP dataset, we use 1-Wasserstein Distance to measure alignment on ordinal response options. **Top:** alignment metric (lower is better) for each aggregated simulation specification, mean across the respective runs. **Bottom:** simulation specification—Survey Response Generation Method, response option variant, model size, and decoding strategy—sorted by the respective alignment metric. Specifications that lead to more than 10% invalid responses are excluded.

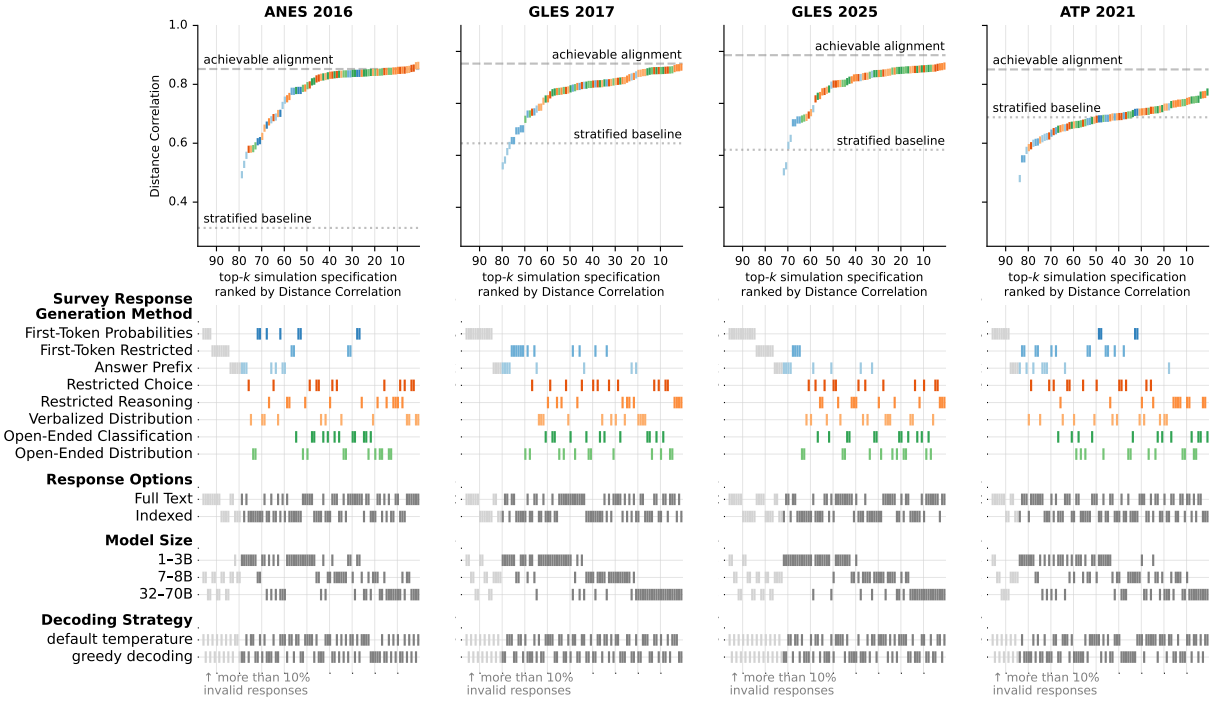
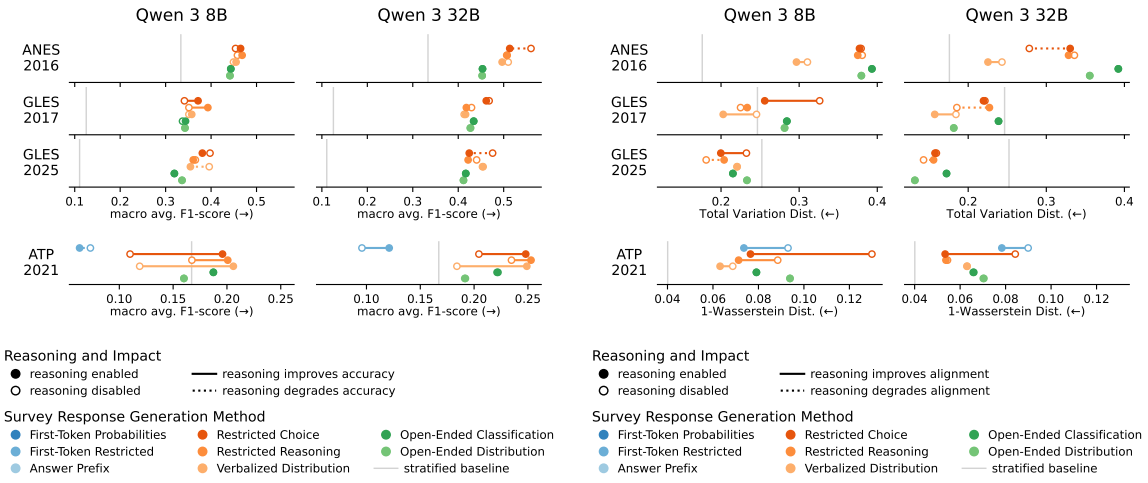


Figure 5: **Subpopulation-Level Alignment—Global Perspective: Distance Correlation. Top:** Distance correlation (higher is better) for each aggregated simulation specification, mean across the respective runs. **Bottom:** simulation specification—Survey Response Generation Method, response option variant, model size, and decoding strategy—sorted by distance correlation. Specifications that lead to more than 10% invalid responses are excluded.



(a) **Individual-Level Alignment.** (b) **Subpopulation-Level Alignment.**

Figure 6: **Alignment With / Without Reasoning.** Mean results shown for default temperature, Full Text response option variants. Methods with more than 1% invalid responses are excluded. **Reasoning does not consistently improve alignment.**

		ANES 2016	GLES 2017	GLES 2025	ATP 2021
Intercept		.343**	.037*	.050	.107**
Survey Response Generation Method	**First-Token Restricted**	.082**	.051*	.066	-.032**
	Answer Prefix	.013	.069**	.059	-.021*
	Restricted Choice	.148**	.242**	.218**	.015
	Restricted Reasoning	.138**	.219**	.196**	.052**
	Verbalized Distribution	.118**	.233**	.216**	.011
	Open-Ended Classif.	.127**	.215**	.184*	.037**
	Open-Ended Distrib.	.114**	.212**	.177*	.018*
Response Option Variant	**Full Text, Reversed**	-.004	.011	.002	-.000
	Indexed	-.042**	-.020	-.009	.006
	Indexed, Reversed	-.054**	-.016	-.016	.005
Model	**Llama 3B**	-.020	.007	-.008	.005
	Llama 70B	-.003	.218**	.214**	.023*
	OLMo 1B	-.073**	-.080**	-.086**	-.040**
	OLMo 7B	.027	.008	-.006	.022*
	OLMo 32B	.061**	.155**	.179**	.034**
	Qwen 8B	-.018	.105**	.128**	.014
	Qwen 8B (R)	-.018	.075**	.107**	.042**
	Qwen 32B	.083**	.218**	.227**	.055**
	Qwen 32B (R)	.038	.160**	.196**	.091**

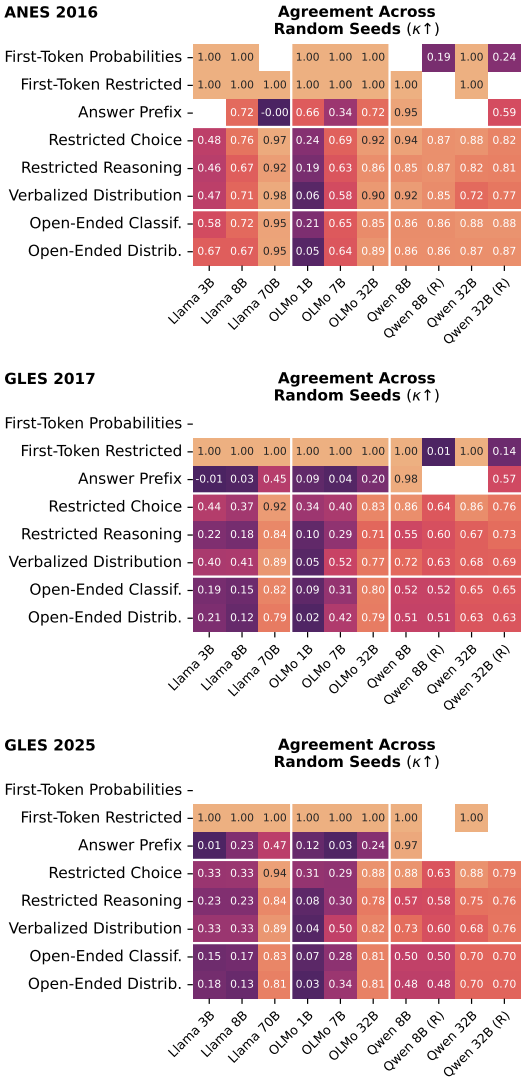
Table 12: **Regression Coefficients for Individual-Level Alignment (↑).** OLS regression per dataset with macro avg. F1-score as the dependent variable (higher is better). We use Survey Response Generation Method, response option variant, and LLM name as independent variables. Reference: First-Token Probabilities Method + Full Text response options + Llama 8B. We do not include interactions to mitigate multicollinearity, and use clustered standard errors across seeds × decoding strategies to appropriately reflect the repeated-measures structure of our evaluation. * $p < 0.05$, ** $p < 0.01$, Benjamini-Hochberg adjusted.

		ANES 2016	GLES 2017	GLES 2025	ATP 2021
Intercept		.322**	.476**	.578**	.070**
Survey Response Generation Method	**First-Token Restricted**	.105**	.044	-.069	.008
	Answer Prefix	.049	-.120**	-.202**	.002
	Restricted Choice	.045	-.165**	-.287**	.012*
	Restricted Reasoning	.061*	-.197**	-.312**	-.010*
	Verbalized Distribution	-.028	-.219**	-.296**	-.016**
	Open-Ended Classif.	.072**	-.174**	-.306**	-.001
	Open-Ended Distrib.	.038	-.216**	-.319**	-.006
Response Option Variants	**Full Text, Reversed**	.003	-.006	.038*	.001
	Indexed	.011	.003	-.001	.002
	Indexed, Reversed	.037**	.012	.026	.002
Model	**Llama 3B**	-.049*	.034	.069**	.002
	Llama 70B	-.047*	-.089**	-.133**	.017**
	OLMo 1B	-.022	.108**	.106**	.026**
	OLMo 7B	-.060**	.064**	.069**	.009*
	OLMo 32B	-.064**	-.073**	-.116**	.014**
	Qwen 8B	.032	.020	-.056**	.022**
	Qwen 8B (R)	-.010	-.013	-.024	.013*
	Qwen 32B	-.070**	-.112**	-.174**	.006*
	Qwen 32B (R)	-.049**	-.069**	-.094**	-.005

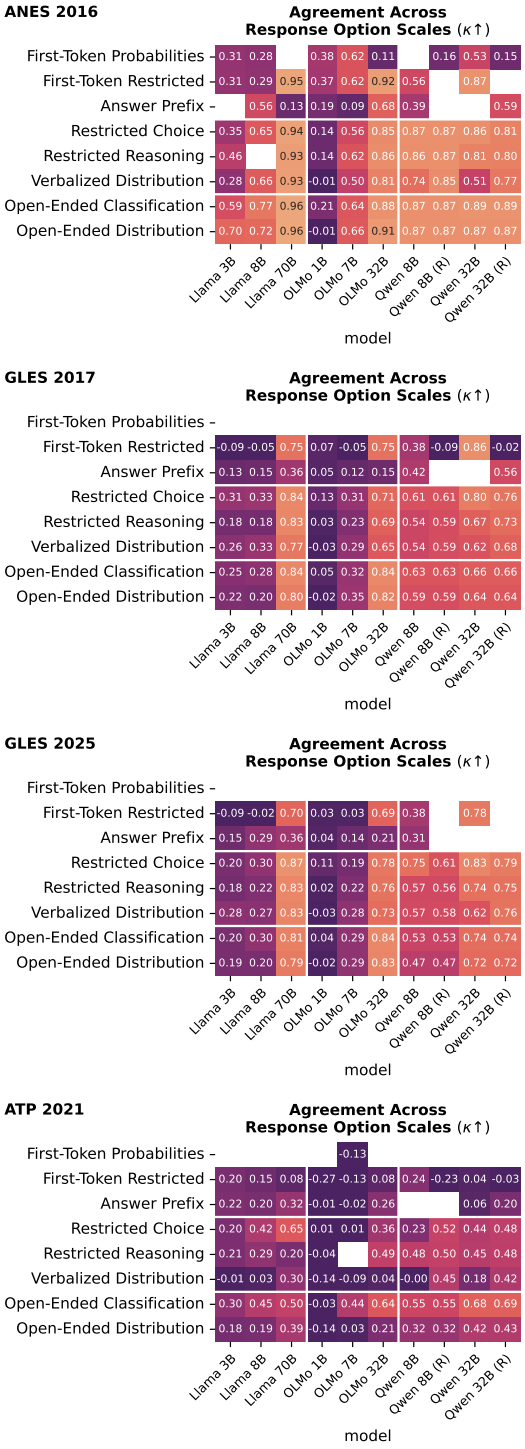
Table 13: **Regression Coefficients for Subpopulation-Level Alignment** ($\downarrow$)**.** OLS regression per dataset with total variation distance (lower is better) as the dependent variable for the ANES and GLES datasets and 1-Wasserstein distance (lower is better) as the dependent variable for the ATP 2021 dataset. We use Survey Response Generation Method, response option variant, and LLM name as independent variables. Reference: First-Token Probabilities Method + Full Text response options + Llama 8B. We do not include interactions to mitigate multicollinearity, and use clustered standard errors across seeds $\times$ decoding strategies to appropriately reflect the repeated-measures structure of our evaluation. * $p < 0.05$, ** $p < 0.01$, Benjamini-Hochberg adjusted.

		Individual-Level		Subpopulation-Level	
		Accuracy (Macro Avg. F1-Score)	**Robustness** (Fleiss' κ Across Scales)	**Alignment** (Total Var./ 1-Wasserst.)	**Global Align.** (Distance Correlation)
Intercept		-1.647**	-1.183**	-0.930**	-1.150**
Survey Response Generation Method	**First-Token Restricted**	0.751**	0.461*	-0.029	0.613*
	Answer Prefix	0.415	0.009	0.667*	-0.073
	Restricted Choice	1.535**	1.178**	0.845**	1.138**
	Restricted Reasoning	1.576**	1.097**	1.090**	1.302**
	Verbalized Distribution	1.443**	0.655**	1.433**	1.104**
	Open-Ended Classif.	1.440**	1.382**	0.882**	1.281**
	Open-Ended Distrib.	1.273**	0.930**	1.097**	1.104**
Dataset	**ATP 2021**	0.035	0.039	0.010	0.037
	GLES 2017	0.002	-0.000	-0.002	0.004
	GLES 2025	0.036	0.038	0.030	0.029
Response Option Variants	**Full Text, Reversed**	-0.034	-0.046	-0.083*	0.049
	Indexed	-0.148**	-0.062*	-0.112**	-0.098**
	Indexed, Reversed	-0.175**	-0.058*	-0.263**	-0.085*
Model	**Llama 3B**	-0.145*	-0.477**	-0.113	-0.404**
	Llama 70B	0.850**	0.971**	0.441**	0.341**
	OLMo 1B	-0.746**	-1.306**	-0.541**	-0.957**
	OLMo 7B	0.175**	-0.388**	-0.063	-0.010
	OLMo 32B	0.850**	0.851**	0.446**	0.585**
	Qwen 8B	0.476**	0.560**	-0.170**	0.278**
	Qwen 8B (R)	0.663**	0.817**	0.171**	0.437**
	Qwen 32B	1.103**	0.916**	0.703**	0.594**
	Qwen 32B (R)	1.163**	1.069**	0.623**	0.679**

Table 14: **Regression Coefficients For Evaluation Metrics (normalized, ↑).** OLS regression on evaluation outcomes for individual-level accuracy, and robustness (Fleiss' κ), as well as subpopulation-level alignment ($1-$ total variation distance / 1-Wasserstein distance) and global alignment (distance correlation). Results from each metric are z-score normalized separately on each dataset. We use dataset, Survey Response Generation Method, response option variant, and LLM as independent variables, and do not include interactions. Reference: ANES 2016 + First-Token Probabilities ■ + Full Text response options + Llama 8B. Results with more than 10% invalid values were excluded. **Restricted Generation Methods ■■■ consistently lead to significant improvements** * $p < 0.05$, ** $p < 0.01$, Benjamini-Hochberg adjusted.

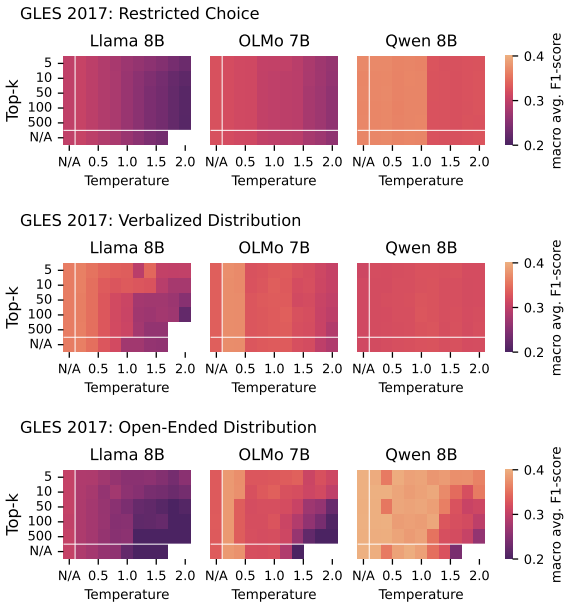


(a) **Individual-Level Agreement Across Seeds.** The First-Token Probability Method and the First-Token Restricted Method achieve perfect agreement across all non-reasoning models, as the output probabilities of the first token are deterministic given the same prompt. No agreement across seeds is calculated for the ATP 2021 dataset, as we evaluated this dataset with only 1 seed to save computational resources.

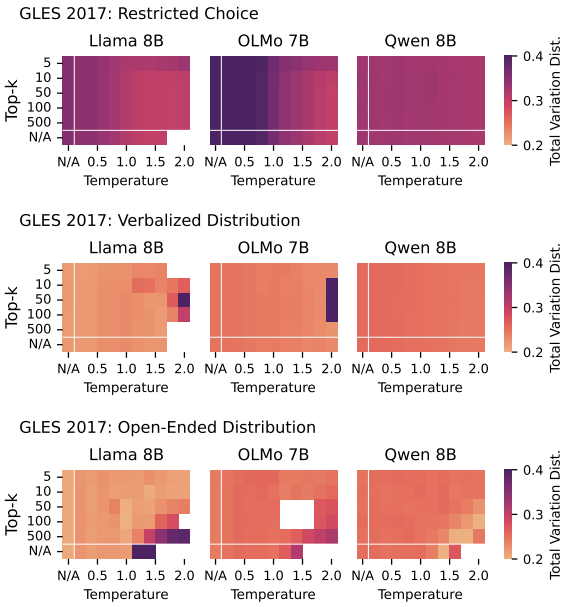


(b) **Individual-Level Agreement Across Response Option Scales.** A sufficient agreement is often desirable, as perturbations in the response options scales should not impact the response that is generated by a model.

Figure 7: **Individual-Level Agreement.** Mean Fleiss's κ (↑), results with more than 10% invalid values excluded. Perfect agreement across seeds or scales might not be considered desirable, as it indicates overly confident individual-level predictions given the variance in human survey responses. For an evaluation of individual-level calibration, see also Figure 9.



(a) **Decoding Strategy Impacts Individual-Level Alignment.** Macro avg. F1-score (↑)—results generally degrade with increasing temperature.



(b) **Decoding Strategy Impacts Subpopulation-Level Alignment.** Total variation distance (↓)—results improve for `Llama 8B` and `OLMo 7B` with the Restricted Choice Method, but are otherwise mostly stable.

Figure 8: **Decoding Strategies.** For 3 Survey Response Generation Methods (Restricted Choice, Verbalized Distribution, and Open-Ended Distribution) and 3 medium-size LLMs (`Llama 8B`, `OLMo 7B`, and `Qwen 8B`), we investigate a diverse range of `temperature` and `top-k` values during decoding. `N/A` stands for greedy decoding or full vocabulary respectively. Results with more than 10% invalid values excluded.

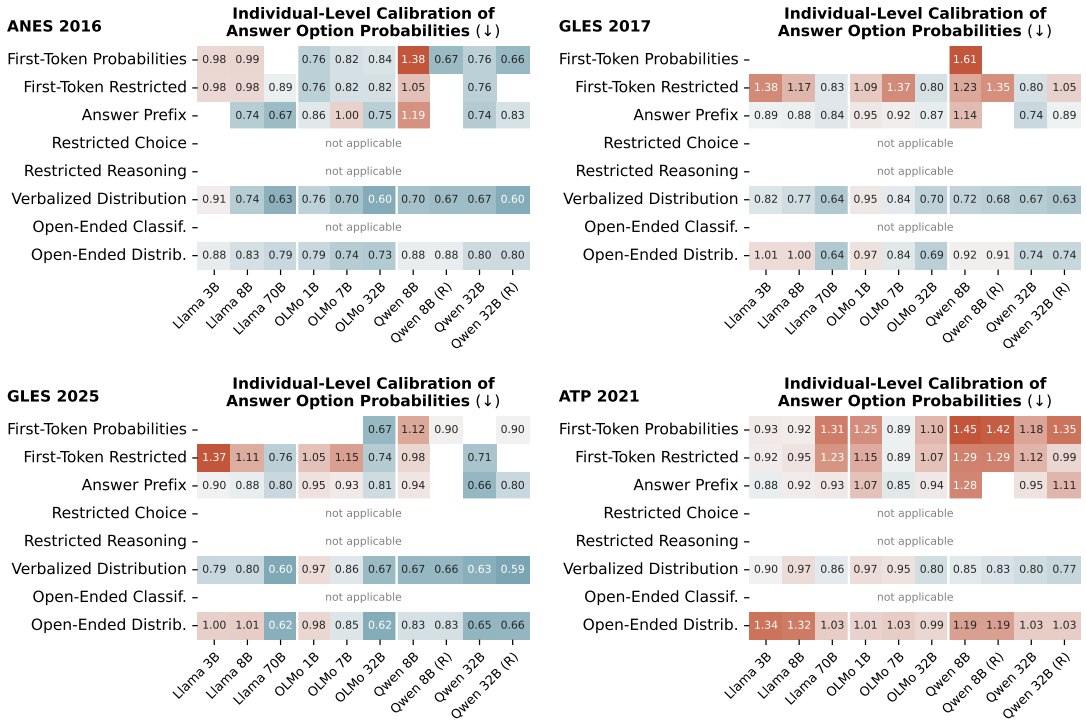


Figure 9: **Individual-Level Calibration** (↓)**.** Mean Brier score (↓) across all response options, results with more than 10% invalid values excluded. For Survey Response Generation Methods that generate individual-level distributions across response options, we can evaluate whether high "confidence" of a model in a response option corresponds with correct predictions on the individual level. The Brier score calculates individual-level calibration as the mean squared error between the confidences and the one-hot encoded human survey responses. Well-calibrated Survey Response Generation Methods are desirable, as they accurately capture individual-level prediction uncertainty. We find that larger models are generally better calibrated than smaller ones. Token Probability-Based Methods can be poorly calibrated for most models, while the **Verbalized Distribution Method leads to the best individual-level calibration.**

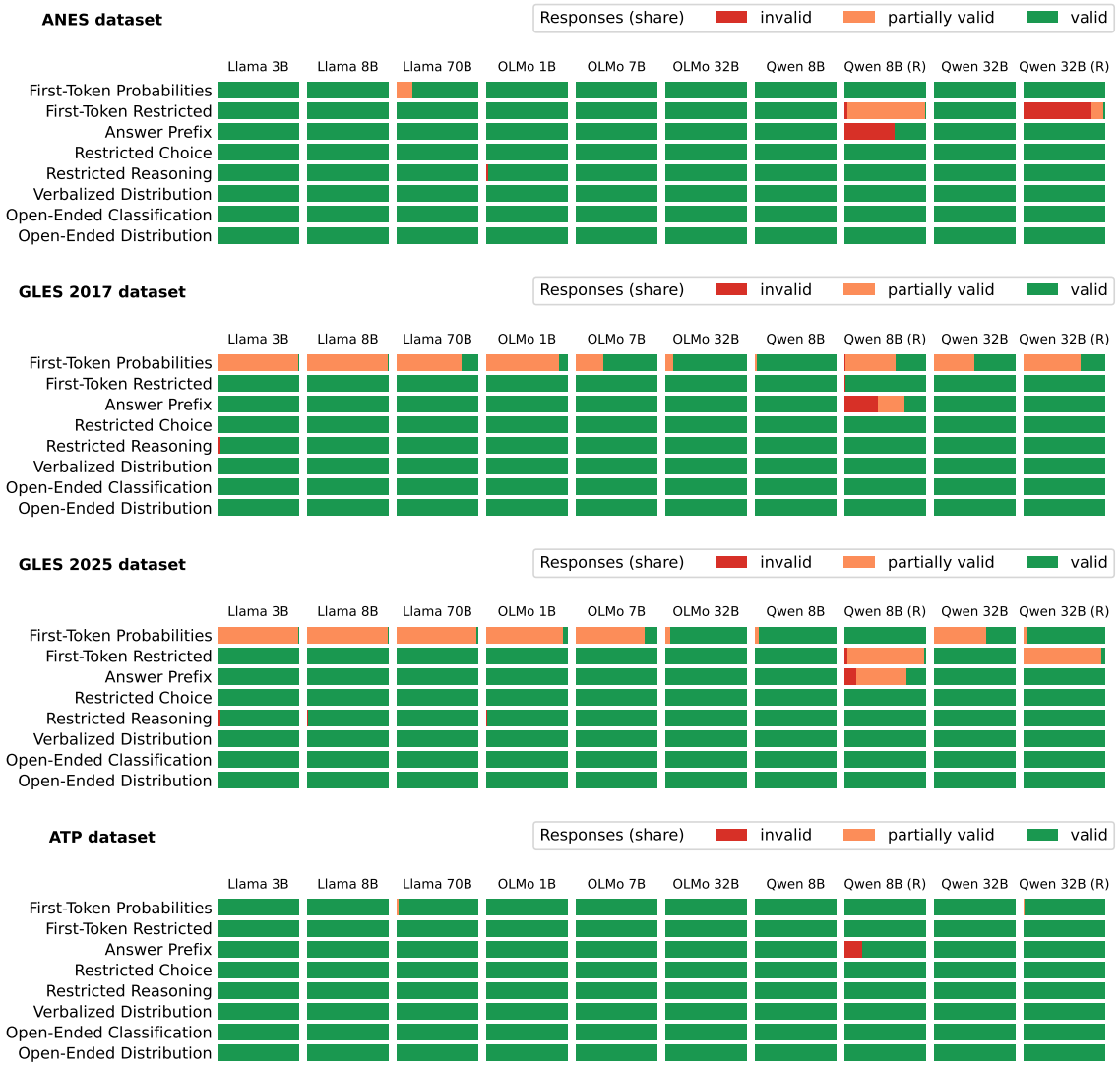


Figure 10: **Fractions of LLM Responses That Are Invalid.** We count model output that does not contain any of the possible closed-ended response options as *invalid.* For Survey Response Generation Methods that generate distributions over possible response options (see Table 1), we count model output that only contains probabilities for some of the response options as *partially valid.*

Survey Response Generation Method	Formatting Instructions (System Prompt)
Token Probability -Based Methods	`You are a political scientist predicting responses to the following question: {QUESTION}. These are the possible answer options: {RESPONSE OPTIONS}. You only respond with the most probable answer option.`
Restricted Choice	`You are a political scientist predicting responses to the following question: {QUESTION}. These are the possible answer options: {RESPONSE OPTIONS}. You only respond with the most probable answer option in the following JSON format:` ```` ```json ```` `{` `"answer": {RESPONSE OPTIONS}` `}` ```` ``` ````
Restricted Reasoning	`You are a political scientist predicting responses to the following question: {QUESTION}. These are the possible answer options: {RESPONSE OPTIONS}. You always reason about the possible answer options first.` `You respond with your reasoning and the most probable answer option in the following JSON format:` ```` ```json ```` `{` `"reasoning": <your reasoning about the answer options>,` `"answer": <{RESPONSE OPTIONS}>` `}` ```` ``` ````
Verbalized Distribution	`You are a political scientist predicting responses to the following question: {QUESTION}. These are the possible answer options: {RESPONSE OPTIONS}. You only respond with a probability for each answer option in the following JSON format:` ```` ```json ```` `{` `{RESPONSE OPTION 1}: <probability>,` `{RESPONSE OPTION 2}: <probability>,` `{...}` `}` ```` ``` ````
Open Generation Methods (Step 1: Open Generation)	`You are a political scientist predicting responses to the following question: {QUESTION}.`
Open-Ended Classification (Step 2: Classification[1])	`You are an expert annotator. These are the possible labels: {RESPONSE OPTIONS}. You only respond with the most probable label in the following JSON format:` ```` ```json ```` `{` `"answer": <{RESPONSE OPTIONS}>` `}` ```` ``` ````
Open-Ended Distribution (Step 2: Classification[1])	`You are an expert annotator. These are the possible labels: {RESPONSE OPTIONS}. You only respond with a probability for each answer option in the following JSON format:` ```` ```json ```` `{` `{RESPONSE OPTION 1}: <probability>,` `{RESPONSE OPTION 2}: <probability>,` `{...}` `}` ```` ``` ````

Table 15: **Formatting Instructions (System Prompts) for English—ANES 2016 & ATP 2021 Datasets.** The `QUESTION` and `RESPONSE OPTIONS` are identical to what was used by Argyle et al. (2023) and Santurkar et al. (2023) respectively. [1]Additional user prompts for the classification step are shown in Table 16.

Survey Response Generation Method	Classification Instructions (User Prompt)
Open-Ended Classification / Open-Ended Distribution (Step 2: Classification)	`Below is a STATEMENT that responds to the question: {QUESTION} Does the following STATEMENT express preference for exactly one of the following answer options: <{RESPONSE OPTIONS}>? If yes, which option is preferred?` `STATEMENT: {OPEN-ENDED MODEL RESPONSE}`

Table 16: **Classification Instructions (User Prompts) for English—ANES 2016 & ATP 2021 Datasets.** These prompts are only used in the second, classification step of the Open-Ended Classification and the Open-Ended Distribution Methods. The user prompts for all other Survey Response Generation Methods, as well as for the first, open-ended step of the Open Generation Methods are shown in Table 19.

Survey Response Generation Method	Formatting Instructions (System Prompt)
Token Probability -Based Methods	Du bist ein Politikwissenschaftler, der Antworten auf die folgende Frage vorhersagt: {QUESTION}. Dies sind die möglichen Antwortoptionen: {RESPONSE OPTIONS}. You only respond with the most probable answer option.
Restricted Choice	Du bist ein Politikwissenschaftler, der Antworten auf die folgende Frage vorhersagt: {QUESTION}. Dies sind die möglichen Antwortoptionen: {RESPONSE OPTIONS}. Du antwortest ausschließlich mit der wahrscheinlichsten Antwortoption im folgenden JSON-Format: ```json { "antwort": {RESPONSE OPTIONS} } ```
Restricted Reasoning	Du bist ein Politikwissenschaftler, der Antworten auf die folgende Frage vorhersagt: {QUESTION}. Dies sind die möglichen Antwortoptionen: {RESPONSE OPTIONS}. Du argumentierst immer zuerst über die möglichen Antwort-Optionen. Du antwortest mit deiner Argumentation und der wahrscheinlichsten Antwort-Option im folgenden JSON-Format: ```json { "argumentation": <deine Argumentation über die Antwort-Optionen>, "antwort": <{RESPONSE OPTIONS}> } ```
Verbalized Distribution	Du bist ein Politikwissenschaftler, der Antworten auf die folgende Frage vorhersagt: {QUESTION}. Dies sind die möglichen Antwortoptionen: {RESPONSE OPTIONS}. Du antwortest ausschließlich mit einer Wahrscheinlichkeit für jede Antwort-Option im folgenden JSON-Format: ```json { {RESPONSE OPTION 1}: <Wahrscheinlichkeit>, {RESPONSE OPTION 2}: <Wahrscheinlichkeit>, {...} } ```
Open Generation Methods (Step 1: Open Generation)	Du bist ein Politikwissenschaftler, der Antworten auf die folgende Frage vorhersagt: {QUESTION}.
Open-Ended Classification (Step 2: Classification[1])	Du bist ein erfahrener Annotator. Das sind die möglichen Labels: {RESPONSE OPTIONS}. Du antwortest nur mit dem wahrscheinlichsten Label im folgenden JSON-Format: ```json { "antwort": <{RESPONSE OPTIONS}> } ```
Open-Ended Distribution (Step 2: Classification[1])	Du bist ein erfahrener Annotator. Das sind die möglichen Labels: {RESPONSE OPTIONS}. Du antwortest nur mit einer Wahrscheinlichkeit für jede Antwortoption im folgenden JSON-Format: ```json { {RESPONSE OPTION 1}: <Wahrscheinlichkeit>, {RESPONSE OPTION 2}: <Wahrscheinlichkeit>, {...} } ```

Table 17: **Formatting Instructions (System Prompts) for German—GLES 2017 & GLES 2025 Datasets.** The QUESTION and RESPONSE OPTIONS are identical to what was used by Von Der Heyde et al. (2025). [1]Additional user prompts for the classification step are shown in Table 18.

Survey Response Generation Method	Classification Instructions (User Prompt)
Open-Ended Classification / Open-Ended Distribution (Step 2: Classification)	Nachfolgend findest du eine AUSSAGE, die auf die Frage {QUESTION} antwortet. Drückt die folgende AUSSAGE eine Präferenz für genau eine der folgenden Antwortoptionen aus: <{RESPONSE OPTIONS}>? Wenn ja, welche Option wird bevorzugt? AUSSAGE: {OPEN-ENDED MODEL RESPONSE}

Table 18: **Classification Instructions (User Prompts) for German—GLES 2017 & GLES 2025 Datasets.** These prompts are only used in the second, classification step of the Open-Ended Classification and the Open-Ended Distribution Methods. The user prompts for all other Survey Response Generation Methods, as well as for the first, open-ended step of the Open Generation Methods are shown in Table 19.

Dataset	Persona & Question (User Prompt)
ANES 2016 (English)	Racially, I am {row.race}. {"I like to discuss politics with my family and friends." if row.discuss_politics == 'Yes' else "I never discuss politics with my family or friends."} Ideologically, I am {row.ideology}. Politically, I am an {row.party}. I {row.church_goer}. I am {int(row.age)} years old. I am a {row.gender}. I am {row.political_interest} interested in politics. It makes me feel {row.patriotism} to see the American flag. I am from {row.state}. In the 2016 presidential election, I voted for
GLES 2017 (German)	Ich bin {row.Alter} Jahre alt und {row.Geschlecht}. Ich habe {row.Bildung}, ein {row.Haushaltseinkommen} monatliches Haushalts-Nettoeinkommen und bin {row['Erwerbstätigkeit']}. Ich bin {row['Religiosität']}. Politisch-ideologisch ordne ich mich {row['Links-Rechts-Einstufung']} ein. Ich identifiziere mich {row['Parteiidentifikation Stärke']} {row['Parteiidentifikation']}. Ich lebe in {row['Ost-West']}. Ich finde, die Regierung sollte die Einwanderung {row['Zuwanderung']} und {row['Einkommensunterschiede verringern']} ergreifen, um die Einkommensunterschiede zu verringern. Habe ich bei der Bundestagswahl 2017 gewählt und wenn ja, welcher Partei habe ich meine Zweitstimme gegeben?
GLES 2025 (German)	– same as for 2017 with only the year in the question changed –
ATP 2021 (English)	In politics today, I consider myself a {row.POLPARTY}. I would describe my political views as {row.POLIDEOLOGY}. My present religion is {row.RELIG}. I am {row.RACE}. The highest level of education I have completed is: {row.EDUCATION}. Last year, my total family income from all sources, before taxes was {row.INCOME}. I currently reside in the {row.CREGION}. I identify as {row.SEX}. {QUESTION}

Table 19: **Persona & Question Prompts (User Prompts) for All Datasets.** All prompts closely follow the templates provided by Argyle et al. (2023); Von Der Heyde et al. (2025); Santurkar et al. (2023). During the evaluations, we omit sentences with missing data on at least one of the variables.