QSTN: A Modular Framework for Robust Questionnaire Inference with Large Language Models

Maximilian Kreutner[1], Jens Rupprecht[1], Georg Ahnert[1], Ahmed Salem[1], Markus Strohmaier[1,2,3]
[1]University of Mannheim, [2]GESIS - Leibniz Institute for the Social Sciences, [3]CSH Vienna

Abstract

We introduce QSTN, an open-source Python framework for systematically generating responses from questionnaire-style prompts to support in-silico surveys and annotation tasks with large language models (LLMs). QSTN enables robust evaluation of questionnaire presentation, prompt perturbations, and response generation methods. Our extensive evaluation (>40 million survey responses) shows that question structure and response generation methods have a significant impact on the alignment of generated survey responses with human answers. We also find that answers can be obtained for a fraction of the compute cost, by changing the presentation method. In addition, we offer a no-code user interface that allows researchers to set up robust experiments with LLMs *without coding knowledge*. We hope that QSTN will support the reproducibility and reliability of LLM-based research in the future.

1 Introduction

Questionnaires have become an important format to probe, assess, and utilize large language models (LLMs) via prompts. Questionnaire-like prompts have been a popular way to evaluate LLMs on tasks such as common knowledge understanding (Hendrycks et al., 2021), language comprehension (Hu et al., 2023; Sravanthi et al., 2024; Kim et al., 2024), and mathematical reasoning (Satpute et al., 2024; Wei et al., 2023). Other work uses existing questionnaires to evaluate LLMs' values; for example, political bias (Röttger et al., 2024; Rozado, 2024), personality traits (Jiang et al., 2024; Shu et al., 2024; Pellert et al., 2024), or psychometric profiles (Ye et al., 2025). With the increasing capability of LLMs, researchers have found additional use cases, such as the creation of synthetic survey responses (Argyle et al., 2023; Ma et al., 2024) or data annotation (Tan et al., 2024).

Despite the widespread use of questionnaire-like prompts, concerns have been raised about the robustness of LLM responses to such prompts. The closed-ended responses of an LLM can vary strongly from its open-ended responses (Röttger et al., 2024; Wang et al., 2024), LLM responses can be biased towards specific survey response options (Tjuatja et al., 2024; Rupprecht et al., 2025), and downstream performance is strongly affected by small changes in the questionnaire configuration (Cummins, 2025; Ahnert et al., 2025).

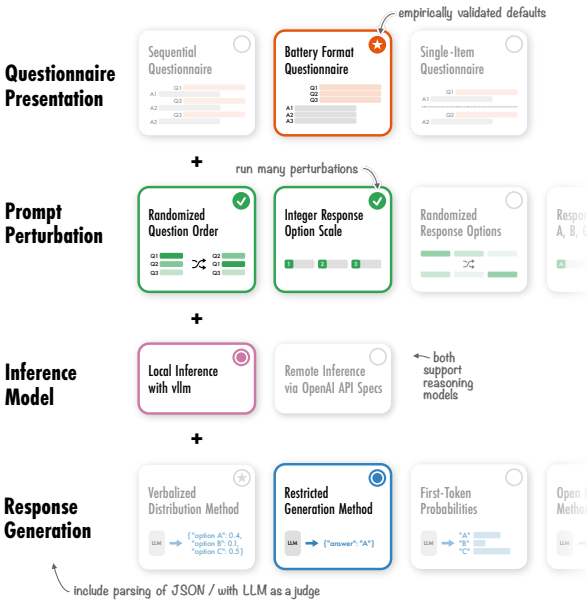

Figure 1: **QSTN Facilitates Easy To Customize and Robust Questionnaire Inference with LLMs.** QSTN provides a fully modular pipeline with different ways to present the questionnaire, prompt perturbations and to choose a response generation method, with automatic parsing. Both local and remote inference are supported.

To address and investigate some of these concerns, **we present QSTN** (pronounced "Question") - a Python framework designed to **facilitate the execution of questionnaire-style experiments with LLMs.** QSTN simplifies the process of creating robust variations of question prompts and answer generation methods, thereby facilitating reproducibility and the analysis of the reliability of LLM-based

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 3: System Demonstrations, pages 537–549

questionnaire research. QSTN provides a complete, modular pipeline, as depicted in Figure 1, for creating the questionnaire presentation, adjusting various parts of the prompt with perturbations, choosing the response generation method, performing inference, and finally, parsing the generated text. We evaluate our framework on more than 40 million survey responses and find that the controlled variation of the experiment pipeline can increase the alignment of generated responses with human survey answers and that the responses can be obtained for a fraction of the compute cost.

Python package under MIT license: https://github.com/dess-mannheim/QSTN

Live GUI: https://hf.co/spaces/qstn/qstn_gui or run it locally by cloning the Git repository

Video: https://youtu.be/uM5Q-Qmm6nQ

2 Core Features

QSTN was developed with three objectives in mind: First, it enables *robust evaluation* of and with LLMs, addressing prompt sensitivity (Tjuatja et al., 2024; Dominguez-Olmedo et al., 2024). QSTN is engineered to address this challenge directly through a highly modular and configurable design. Each part of the pipeline can be exchanged independently from the other parts.

Second, QSTN is designed to be *efficient*, so it can be used in large-scale studies. For experiments with multiple prompt variations and/or personas, we automatically utilize prefix caching and batching for local inference in vLLM (Kwon et al., 2023), and asynchronous calling with the AsyncOpenAI API (OpenAI, 2023).

Finally, QSTN is designed to be as *easy to use as possible*. Since we maintain the common prompt format of the system prompt and user prompt, adapting a project to QSTN is seamless. The package offers a complete pipeline from prompt creation and inference to parsing, which can be done in only three function calls to the package. Integration with existing vLLM and OpenAI packages is straightforward.

QSTN's *core strength lies in its ability to systematically and easily control and vary the setup* of questionnaire-like prompting experiments. The following aspects of the experiments can be exchanged and varied by simply switching out one module for another.

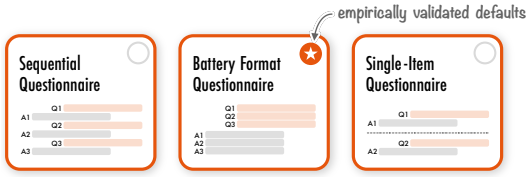

Figure 2: **QSTN Questionnaire Presentation Modes**

2.1 Questionnaire Presentation

QSTN supports three distinct questionnaire presentation modes, as shown in Figure 2:

- **Sequential:** Each question is asked in the same conversation context in multiple, sequential chat calls.
- **Battery:** All questions are asked in battery and the model is expected to answer all questions in one response in the same context.
- **Single-item:** Each question is asked in a new context, with the LLM not being aware of the previous questions and answers.

Questionnaire presentation is a fundamental decision to make when using LLMs with questionnaire-like prompts. For example, if we want to annotate data, is it better to give all annotation questions in the same prompt, or should each question be asked in a new context? There is evidence that keeping multiple tasks in the prompts can improve variety for creative writing (Zhang et al., 2025) and improve performance for classification tasks in moral foundations (Chen et al., 2025). LLMs are also able to perform multiple tasks of different kinds in battery (Son et al., 2024), which can save computing time.

2.2 Prompt Perturbation

Previous studies found that LLMs synthetic survey responses are highly sensitive to prompt perturbations and exhibit biases, such as token biases, recency bias, or A-bias (Pezeshkpour and Hruschka, 2024; Li and Gao, 2025; Rupprecht et al., 2025; Dominguez-Olmedo et al., 2024; Röttger et al., 2024). QSTN can automatically randomize or reverse both the order of the questions within the survey and the order of answer options for each question to identify and mitigate these biases. This ensures that high performance is robust and independent of ordering. Previous research has found that LLMs can be sensitive to small changes in prompt format (He et al., 2024; Sclar et al., 2024). QSTN allows users to define custom answer label schemas (e.g., A/B/C, 1/2/3, i/ii/iii), enabling rigorous testing of a model's robustness to superficial

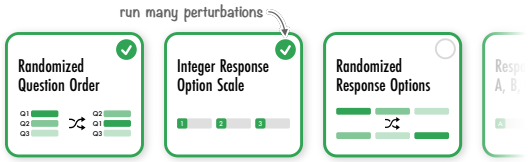

Figure 3: **QSTN Supported Prompt Perturbations**

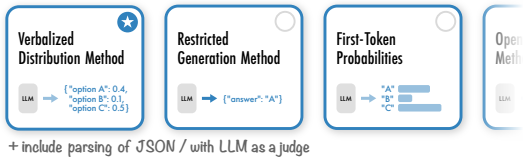

Figure 4: **QSTN Response Generation Methods**

formatting changes. QSTN can perform the following Answer Option Perturbations, which are shown in Figure 3:

- **Reversed Response Order:** The order of answer options is reversed (e.g., a scale from '1: Very important' to '5: Not important' becomes '1: Not important' to '5: Very important').
- **Missing Refusal Option:** The "Don't know" or refusal option is removed from the list of choices.
- **Odd/Even Scale Transformation:** For scales with an even number of options, a semantically appropriate middle category is added, transforming it into an odd-numbered scale (e.g., by adding 'Neutral'). Conversely, for odd-numbered scales, we remove the middle category to create an even scale and adjust the integer label accordingly.

In addition, QSTN can perform the following Question Perturbations:

- **Typographical Errors:** three types of typos can be introduced: *Key Typo* (replacing a character with a random one), *Letter Swap* (swapping two adjacent characters in a random word), and *Keyboard Typo* (replacing a character with an adjacent one on a QWERTY keyboard).
- **Semantic Variations:** Additional semantic variations can be introduced while preserving the original meaning: first, by *Synonym Replacement*, where a variable amount of words in the original question are replaced with synonyms. Second, through *Paraphrasing* the entire question is rephrased.

2.3 ■ Response Generation

While generative language models are designed to generate open-ended text, previous studies have implemented various approaches to constrain LLMs to closed-ended responses (e.g., Ma et al., 2024). We define **Response Generation Methods** as techniques used to elicit closed-ended responses from large language models to questionnaires (Ahnert et al., 2025). QSTN supports the following Response Generation Methods, with examples being shown in Figure 4:

- **Token Probability-Based Methods:** Extract probabilities for response options from the output token probabilities of an LLM.
- **Restricted Generation Methods:** Force the model to respond only with designated response options using formatting instructions in the prompt and (optionally) restrict the vocabulary of the LLM through *structured outputs*.
- **Open Generation Methods:** Generate open-ended responses first and then classify them in a second step.

The Restricted Generation Methods can be used to generate exactly one of the available response options—optionally in JSON format, or with reasoning—or to generate a **verbalized distribution** of probabilities for all response options, following Meister et al. (2025). All Response Generation Methods can be adjusted to, e.g., have the model generate a prefix before token probabilities are extracted. QSTN includes **suitable parsers** for all generated responses: JSON & LLM-as-a-judge.

3 Using QSTN

The package containing QSTN can be installed in the desired environment using pip. We support both a lightweight installation with `pip install qstn`, which only requires dependencies for API usage, and a full installation with `pip install qstn[vllm]`, which allows for local inference with vllm. QSTN is easily integrable into current workflows, requiring just a total of three function calls for the most basic functionality, and it still allows users to freely define their prompts. A minimum usage example is given in Listing 1. By simply exchanging the function in the inference step, the questionnaire presentation can be adjusted, or a different type of parser can be selected. Additionally, building on this simple example, only one more module is needed to implement controlled prompt perturbations and response generation methods.

Non-Code User Interface QSTN offers a User Interface to create and run inference with LLMs without having to program any Python code. The UI offers the same core functionality as the main frame-

work, allowing users to upload questionnaires, systematically alter the prompt structure, set model parameters, and run inference. While the UI generally offers the same functions as the coding package, some more advanced features, such as inferencing models directly through the vllm API, are currently not supported.

4 Evaluation

We evaluate QSTN primarily on the task of generating synthetic survey responses, which is a topic of growing interest. Our results demonstrate that our proposed variations significantly influence both the alignment of synthetic data with real-world responses and computational efficiency. Across all experiments, we use the following instruction-finetuned versions of the models: LLama3 1-70B (Grattafiori et al., 2024), Qwen3 4-30B (Yang et al., 2025), Phi-4-mini (Abdin et al., 2024), Gemma3 4-27B (Team et al., 2025), OLMo2 1-32B (OLMo et al., 2024), Yi1.5 6B (Young et al., 2024), and Gemini1.5 Pro (Team et al., 2024). We present new results and evaluations regarding questionnaire presentation and provide an overview of previous experiments that were conducted in or implemented into QSTN, which evaluate Prompt Perturbations and Response Generation Methods.

```python
import qstn
import pandas as pd
from vllm import LLM

# 1. Prepare questionnaire and persona data
questionnaires = pd.read_csv("hf://datasets/qstn/ex/q.csv")
personas = pd.read_csv("hf://datasets/qstn/ex/p.csv")
prompt = (
    f"Please tell us how you feel about:\n"
    f"{qstn.utilities.placeholder.PROMPT_QUESTIONS}"
)
interviews = [
    qstn.prompt_builder.LLMPrompt(
        questionnaire_source=questionnaires,
        system_prompt=persona,
        prompt=prompt,
    ) for persona in personas.system_prompt]

# 2. Run Inference
model = LLM("Qwen/Qwen3-4B", max_model_len=5000)
results = qstn.survey_manager.conduct_survey_single_item(
    model, interviews, max_tokens=500
)

# 3. Parse Results
parsed_results = qstn.parser.raw_responses(results)
```

Listing 1: **Minimum usage example of QSTN.** QSTN can be easily integrated into existing projects, requiring just three function calls to operate. Users familiar with vllm or the OpenAI API can use the same Model/Client calls and arguments. In this example reasoning and the generated response are automatically parsed.

4.1 Questionnaire Presentation

We start by demonstrating that the presentation of the questionnaire significantly impacts the subpopulation alignment of generated responses with real answers. Furthermore, selecting the optimal method results in savings for both token usage and GPU time. We test three fundamentally different presentations, as described in Section 2.1.

We base our experiments on Bisbee et al. (2024), where respondents of the ANES survey are instructed to consider a certain group and to indicate the degree to which they experience warm (positive, affectionate, etc.) or cool (negative, disdainful, etc.) feelings toward members of that group on a scale from 0 to 100. For each of the 7530 participants, we use three different seeds, which leads to a **total of 10,843,200 individual question responses** across 16 questions, 10 different models, and 3 different presentations.

We use the same prompts as in the initial study, with the addition of an instruction on how to format the output to align with the response generation method. Our full prompts can be seen in Table 7 in the Appendix. Respondents were stratified into subpopulations based on the intersection of gender, race, and ideology (see Appendix Table 6 for full subpopulation attributes). We measure individual alignment via Mean Absolute Error and subpopulation distributional alignment via Wasserstein distance; results are displayed in Table 1. To quantify the effects of questionnaire presentation, we fitted Ordinary Least Squares and Weighted Least Squares models for MAE and Wasserstein distance, respectively. Both models include interaction terms between presentation and model, as well as fixed effects for iteration seeds. The ■ single-item presentation and `Llama-3.3-70B-Instruct` serve as the reference categories.

We find that questionnaire presentation has a substantial impact on distributional alignment, whereas the effects on individual-level accuracy, while statistically significant, are practically marginal. For the reference model, the ■ battery presentation yields the strongest improvement in subpopulation alignment ($\beta_{\text{WD}} = -1.17, p < 0.01$), representing an approximate 8% better alignment than with the ■ single-item presentation. This effect is consistent across the large models we tested, as the interaction effect for both `Qwen-30B` and `Gemma-27B` was not statistically significant. However, for smaller models, the effect is highly

questionnaire presentation	Mean Absolute Error ↓ ■ sequential	■ battery	■ single-item	Wasserstein distance ↓ ■ sequential	■ battery	■ single-item
gemma-3-4b-it	20.96 ± 0.02	21.92 ± 0.01	**19.94 ± 0.02**	16.48 ± 0.01	17.62 ± 0.02	**16.39 ± 0.01**
gemma-3-12b-it	18.26 ± 0.01	**18.07 ± 0.02**	19.11 ± 0.01	14.53 ± 0.02	**13.44 ± 0.01**	16.44 ± 0.01
gemma-3-27b-it	**17.59 ± 0.01**	17.90 ± 0.01	18.01 ± 0.01	**14.00 ± 0.01**	14.26 ± 0.01	15.17 ± 0.00
Llama-3.2-1B-Instruct	30.89 ± 0.25	**30.22 ± 0.07**	35.69 ± 0.12	18.66 ± 0.33	**18.15 ± 0.12**	27.52 ± 0.17
Llama-3.2-3B-Instruct	24.20 ± 0.10	**22.98 ± 0.04**	24.32 ± 0.06	**13.14 ± 0.11**	13.50 ± 0.03	15.88 ± 0.07
Llama-3.1-8B-Instruct	21.01 ± 0.04	20.88 ± 0.02	**20.87 ± 0.04**	13.62 ± 0.02	**12.90 ± 0.02**	14.11 ± 0.04
Llama-3.3-70B-Instruct	18.23 ± 0.00	**17.67 ± 0.00**	17.87 ± 0.01	14.18 ± 0.00	**13.56 ± 0.01**	14.73 ± 0.01
Phi-4-mini-instruct	20.98 ± 0.03	**19.72 ± 0.01**	21.23 ± 0.03	**11.69 ± 0.06**	12.21 ± 0.02	14.56 ± 0.05
Qwen3-4B-Instruct-2507	**19.29 ± 0.02**	20.34 ± 0.01	20.05 ± 0.02	**13.75 ± 0.02**	15.59 ± 0.01	15.18 ± 0.01
Qwen3-30B-A3B-Instruct-2507	17.68 ± 0.02	**17.67 ± 0.01**	18.29 ± 0.01	13.88 ± 0.02	**13.68 ± 0.01**	15.21 ± 0.02

Table 1: **Individual and subpopulation alignment based on ■ questionnaire presentation.** Mean absolute error for each individual response and weighted mean Wasserstein distance across the subpopulations. Wasserstein distance significantly improves with sequential and battery presentation for most models, compared to single-item.

architecture-dependent: `Phi-4-mini` achieves the best overall alignment in our experiment using the ■ sequential presentation, whereas `gemma-3-4b` achieves the best alignment with ■ single-item presentation.

Considering the large differences in tokens and compute time between the presentation methods (shown in Table 2), **we recommend the ■ battery presentation as the default for future questionnaire-based experiments with large persona prompts**. However, thorough tests should be conducted to ensure that performance is comparable to other presentations for the specific model and task at hand. `QSTN` makes these validation experiments accessible by requiring just a single method change in the pipeline.

Presentation	Calls	Input T.	Output T.	Inference Time
■ sequential	16	8216	288	09:29:05
■ battery	1	723	142	01:34:45
■ single-item	16	4288	288	03:22:23

Table 2: **API Calls, Tokens and inference time of different ■ questionnaire presentations.** We report the number of API calls, tokens and inference time for the largest model LLama-3.3-70B-Instruct. The tokens are calculated on one persona and the time is measured by a whole run of 7530 personas with 3 seeds. All experiments have been conducted with vllm on two 2 NVIDIA H100 GPUs (tensor-parallel).

4.2 ■ Prompt Perturbation

In previous research (Rupprecht et al., 2025), we found a consistent recency bias in all nine models tested, favoring the same answer option when placed at the end of the options list instead of the beginning. This effect was substantial, with the selection frequency of the semantically same option increasing by more than 20 times for `Llama-3.1-8B` when moved to the last position, while all other configurations, such as question and prompt phrasing, were kept constant.

All models facing prompt perturbations showed some level of non-robust responses, whereas larger models such as `Llama-3.3-70B` and `Gemini-1.5-Pro` respond more robustly. The magnitude of the effect of perturbations (e.g., on the answer option or the question phrasing) on response robustness mainly depends on the type of perturbation applied. We identified that some of the ■ Answer Option Perturbations and ■ Question Perturbations have a larger impact on response robustness than others (see Table 3). Reversing the answer options or introducing typos or paraphrasing the questions is more harmful to robustness than swapping characters within a word or removing the refusal category. In addition, we found that 67% and 89% of models select the middle category significantly more often when a 5- or 11-point Likert scale is provided, respectively.

These findings underline the importance of robustness checks, e.g., through prompt perturbations. `QSTN` allows the user to apply various perturbations automatically to any questionnaire presented and thus assess the response robustness of the LLM.

4.3 ■ Response Generation Methods

To investigate the impact of Response Generation Methods on generated questionnaire responses, we **predict survey responses to questions of political attitudes** in the American National Election

Model	Answer Options (1)	(2)	(3)	Question Perturbations (4)	(5)	(6)	(7)	(8)
Llama-3.3-70B	0.50	0.73	**0.60**	0.52	**0.76**	0.58	0.58	**0.66**
Llama-3.1-8B	0.08	0.39	0.27	0.32	0.31	0.23	0.32	0.16
Llama-3.2-3B	0.10	0.11	0.16	0.10	0.16	0.18	0.23	0.10
Llama-3.2-1B	0.00	0.11	0.03	0.05	0.11	0.00	0.13	0.02
Gemini-1.5-Pro	**0.69**	0.76	0.55	**0.68**	0.73	**0.66**	0.60	0.55
Phi-3.5-mini	0.53	**0.81**	0.45	0.50	0.61	0.47	**0.71**	0.53
Mistral-7B-v0.3	0.68	**0.81**	0.53	0.58	0.65	0.60	**0.71**	0.53
Qwen-2.5-7B	0.32	0.48	0.45	0.48	0.65	0.45	0.55	0.44
Yi-1.5-6B	0.47	0.68	0.55	0.50	0.50	0.45	0.65	0.29

Table 3: **Impact of ■ Answer Option and ■ Question Perturbations on the Response Robustness of different LLMs (↑).** Share of fully robust responses per model. Bold indicates the highest robustness score for that perturbation type. Perturbation Keys: (1) Reversed Answer Options, (2) Missing Refusal, (3) Even Scale, (4) Key Typos, (5) Letter Swap, (6) Keyboard Typos, (7) Synonyms, (8) Paraphrase

Response Generation Method	ANES 2016	GLES 2017	GLES 2025	ATP 2021
Intercept	.374*	.312*	.288*	.503*
■ First-Token Prob.	-.003	.147*	.194*	-.049*
■ Verbalized Distrib.	**-.074***	**-.057***	-.013	**-.168***
■ Open-Ended Distrib.	-.006	-.052*	**-.037***	-.082*

Table 4: **Impact of ■ Response Generation Methods on Subpopulation-Level Alignment (↓).** OLS regression coefficients by dataset with total variation distance (↓) as the dependent variable and Survey Response Generation Method, prompt perturbation, and LLM as independent variables. We show coefficients for selected Response Generation Methods (Reference: Restricted Choice)—see Appendix B for all coefficients and more details on OLS model choice. **The Verbalized Distribution Method leads to significant improvements.** * $p < 0.05$ (Benjamini–Hochberg corrected)

Study (ANES, 2016), the German Longitudinal Election Study (GLES, 2017, 2025), and the American Trends Panel (ATP, 2021). We thereby partially replicate the studies by Argyle et al. (2023), von der Heyde et al. (2025), and Santurkar et al. (2023), while extending them to include additional Response Generation Methods. We compare 8 Response Generation Methods on 10 open-weight LLMs, including reasoning models. For robustness, we include 4 prompt variations, 3 random seeds for temperature-scaled decoding, as well as greedy decoding. Overall, **we simulate 32 mio. survey responses with `QSTN`**, and evaluate their alignment with human survey responses on individual and subpopulations levels. For subpopulation-level alignment, we split the set of respondents into subpopulations by considering all unique values of all persona attributes that were included in the studies we replicate, e.g., women & men, people from different states, etc. We report the subpopulation-level alignment on categorical response distributions using total variation distance (see also Meister et al., 2025; Baan et al., 2022).

Table 4 shows selected OLS regression coefficients for subpopulation-level alignment. We find that the Verbalized Distribution Method yields significant improvements on most datasets. In combination with the individual-level alignment results presented in Ahnert et al. (2025), we conclude that: (i) 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 show significant improvement over other methods while also being more computationally efficient than Open Generation Methods.

5 Related Work

Due to the importance of controlled prompt perturbation, a number of frameworks have started to address this issue. In general, `QSTN` supports controlled variation and combines it with the pipeline to allow for automatic parsing of all prompt variations. Additionally, as `QSTN` allows for modular prompts, these frameworks can be used in conjunction with it. PromptSuite (Habba et al., 2025) focuses on prompt perturbation through paraphrasing and formatting. PromptSource (Bach et al., 2022) is a framework for making and sharing different types of natural language prompts. Prompt-Agnostic Fine-Tuning (PAFT) (Wei et al., 2025) varies prompts in the fine-tuning process rather than during inference.

There are also frameworks that model the entire pipeline of LLM experiments, similar to `QSTN`. Unitxt (Bandel et al., 2024) is an open-source Python framework for data processing pipelines. While powerful, it requires users to understand the Unitxt operator language, which can add cognitive overhead. The EDSL framework (Horton and Horton, 2024) can be used to run surveys with LLMs, but it does not provide full freedom over the exact system prompt or prompt and the Response Generation Method.

6 Conclusion

We introduce QSTN, a Python framework designed to make LLM inference with questionnaires more robust. Our evaluation demonstrates that by enabling controlled variations in the generation process, QSTN can significantly improve the alignment of generated responses with human answers while reducing inference costs. A core feature of QSTN is its modularity, allowing researchers to easily vary their experimental setup with only minimal additional coding effort. The framework is broadly applicable to tasks such as data annotation, synthetic data generation, persona studies, and the analysis of LLM behavior itself.

Limitations

Currently, our evaluation is primarily focused on the creation of synthetic survey responses. We hope that by releasing QSTN to the open-source community, more robust experiments can be conducted in other application domains. While we support a variety of different Response Generation Methods and parsing options, we currently do not support every type of structured output; for example, we do not support output that is guided by a regex pattern or context free grammar. As such, not every type of experiment can currently be conducted in QSTN. We hope that by making the project open-source, we will be able to support more ways to conduct experiments. Additionally, while we plan to add support for non-instruct models, they are currently not supported.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Georg Ahnert, Anna-Carolina Haensch, Barbara Plank, and Markus Strohmaier. 2025. Survey response generation: Generating closed-ended survey responses in-silico with large language models. *arXiv preprint arXiv:2510.11586*.

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.

Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernandez. 2022. Stop measuring calibration when humans disagree. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1892–1915, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Stephen H. Bach, Victor Sanh, Zheng-Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-David, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Alan Fries, and 8 others. 2022. Promptsource: An integrated development environment and repository for natural language prompts. *Preprint*, arXiv:2202.01279.

Elron Bandel, Yotam Perlitz, Elad Venezian, Roni Friedman, Ofir Arviv, Matan Orbach, Shachar Don-Yehiya, Dafna Sheinwald, Ariel Gera, Leshem Choshen, Michal Shmueli-Scheuer, and Yoav Katz. 2024. Unitxt: Flexible, shareable and reusable data preparation and evaluation for generative AI. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations)*, pages 207–215, Mexico City, Mexico. Association for Computational Linguistics.

James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. 2024. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32(4):401–416.

Ziyu Chen, Junfei Sun, Chenxi Li, Tuan Dung Nguyen, Jing Yao, Xiaoyuan Yi, Xing Xie, Chenhao Tan, and Lexing Xie. 2025. MoVa: Towards generalizable classification of human morals and values. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33204–33248, Suzhou, China. 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*.

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. Curran Associates, Inc.

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

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Eliya Habba, Noam Dahan, Gili Lior, and Gabriel Stanovsky. 2025. PromptSuite: A task-agnostic framework for multi-prompt generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 254–263, Suzhou, China. Association for Computational Linguistics.

Jia He, Mukund Rungta, David Koleczek, Arshdeep Sekhon, Franklin X Wang, and Sadid Hasan. 2024. Does prompt formatting have any impact on llm performance? *arXiv preprint arXiv:2411.10541*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

John Horton and Robin Horton. 2024. Edsl: Expected parrot domain specific language for ai powered social science. Whitepaper, Expected Parrot.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, Toronto, Canada. Association for Computational Linguistics.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627, Mexico City, Mexico. Association for Computational Linguistics.

Yeeun Kim, Youngrok Choi, Eunkyung Choi, JinHwan Choi, Hai Jin Park, and Wonseok Hwang. 2024. Developing a pragmatic benchmark for assessing Korean legal language understanding in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5573–5595, Miami, Florida, USA. 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*, SOSP '23, page 611–626, New York, NY, USA. Association for Computing Machinery.

Ruizhe Li and Yanjun Gao. 2025. Anchored answers: Unravelling positional bias in GPT-2's multiple-choice questions. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2439–2465, Vienna, Austria. Association for Computational Linguistics.

Bolei Ma, Xinpeng Wang, Tiancheng Hu, Anna-Carolina Haensch, Michael 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.

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.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, and 1 others. 2024. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*.

OpenAI. 2023. OpenAI Python Library.

Max Pellert, Clemens M. Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. 2024. Ai psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, 19(5):808–826. Epub 2024 Jan 2.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico. Association for Computational Linguistics.

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.

David Rozado. 2024. The political preferences of llms. *Preprint*, arXiv:2402.01789.

Jens Rupprecht, Georg Ahnert, and Markus Strohmaier. 2025. Prompt perturbations reveal human-like biases in llm survey responses. *arXiv preprint arXiv:2507.07188*.

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*, ICML'23. JMLR.org.

Ankit Satpute, Noah Gießing, André Greiner-Petter, Moritz Schubotz, Olaf Teschke, Akiko Aizawa, and Bela Gipp. 2024. Can llms master math? investigating large language models on math stack exchange.

In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 2316–2320.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*.

Bangzhao Shu, Lechen Zhang, Minje Choi, Lavinia Dunagan, Lajanugen Logeswaran, Moontae Lee, Dallas Card, and David Jurgens. 2024. You don't need a personality test to know these models are unreliable: Assessing the reliability of large language models on psychometric instruments. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5263–5281, Mexico City, Mexico. Association for Computational Linguistics.

Guijin Son, SangWon Baek, Sangdae Nam, Ilgyun Jeong, and Seungone Kim. 2024. Multi-task inference: Can large language models follow multiple instructions at once? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5606–5627, Bangkok, Thailand. Association for Computational Linguistics.

Settaluri Sravanthi, Meet Doshi, Pavan Tankala, Rudra Murthy, Raj Dabre, and Pushpak Bhattacharyya. 2024. PUB: A pragmatics understanding benchmark for assessing LLMs' pragmatics capabilities. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12075–12097, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA. Association for Computational Linguistics.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

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*, 0(0):1–23.

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.

Chenxing Wei, Mingwen Ou, Ying He, Yao Shu, and Fei Yu. 2025. PAFT: Prompt-agnostic fine-tuning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 694–717, Suzhou, China. Association for Computational Linguistics.

Tianwen Wei, Jian Luan, Wei Liu, Shuang Dong, and Bin Wang. 2023. Cmath: Can your language model pass chinese elementary school math test? *arXiv preprint arXiv:2306.16636*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Haoran Ye, Jing Jin, Yuhang Xie, Xin Zhang, and Guojie Song. 2025. Large language model psychometrics: A systematic review of evaluation, validation, and enhancement. *Preprint*, arXiv:2505.08245.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, and 1 others. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Jiayi Zhang, Simon Yu, Derek Chong, Anthony Sicilia, Michael R Tomz, Christopher D Manning, and Weiyan Shi. 2025. Verbalized sampling: How to mitigate mode collapse and unlock llm diversity. *arXiv preprint arXiv:2510.01171*.

A Questionnaire Presentation

As another measure of individual evaluation of predictions, we calculate the Pearson correlation between predictions and the ground truth and present it in Table 5. Similarly to the Mean Absolute Error, we see little difference in the performance of the different questionnaire presentations.

We show all attributes we considered for the subpopulation analysis for Wasserstein distance in Table 6. The full regression results for both MAE

questionnaire presentation	■ sequential	■ battery	■ single-item
gemma-3-4b-it	**0.59 ± 0.00**	0.55 ± 0.00	0.57 ± 0.00
gemma-3-12b-it	0.62 ± 0.00	0.62 ± 0.00	**0.64 ± 0.00**
gemma-3-27b-it	**0.62 ± 0.00**	0.61 ± 0.00	0.61 ± 0.00
Llama-3.2-1B-Instruct	**0.25 ± 0.01**	0.18 ± 0.00	0.10 ± 0.00
Llama-3.2-3B-Instruct	0.51 ± 0.00	0.49 ± 0.00	**0.52 ± 0.00**
Llama-3.1-8B-Instruct	0.56 ± 0.00	**0.57 ± 0.00**	0.56 ± 0.00
Llama-3.3-70B-Instruct	**0.64 ± 0.00**	**0.64 ± 0.00**	**0.64 ± 0.00**
Phi-4-mini-instruct	0.48 ± 0.00	0.49 ± 0.00	**0.52 ± 0.00**
Qwen3-4B-Instruct-2507	**0.60 ± 0.00**	0.55 ± 0.00	0.59 ± 0.00
Qwen3-30B-A3B-Instruct-2507	**0.62 ± 0.00**	**0.62 ± 0.00**	0.59 ± 0.00

Table 5: **Mean and Standard Deviation of Pearson Correlation between Prediction and Ground Truth.** Similar to Mean Absolute Error, individual alignment measured with Pearson Correlation shows little difference between different questionnaire presentations.

and Wasserstein Distance can be seen in 8. We report the coefficients and the Benjamini-Hochberg corrected p-values. Additionally, we want to determine if the questionnaire presentation has different effects on different questions. For this, we fit an additional Weighted Least Squares regression on all subpopulations based on the full interaction between the questionnaire presentation, the model, and the specific interview question. We set ■ single-item, the biggest model Llama-3.3-70B-Instruct and the first question as the reference categories, as for this question the LLM has no answers for the other questions in context regardless of the questionnaire presentation.

All questions show improvements, and a subset of five questions shows statistically significant improvement ($p < 0.05$) when using ■ battery presentation instead of ■ single-item presentation. The largest improvement is in the question about feelings towards the group of Gays and Lesbians ($\beta = -3.82, p < 0.01$) when using ■ battery presentation. Figure 5 visually confirms this: when previous questions and answers are included in the context, the model's response distribution aligns much more closely with the ground truth, exhibiting a similar tendency toward neutral answers. The other significant questions concern the groups of White Americans, Asian Americans, Christians, and Liberals.

We use the same prompt as that used in Bisbee et al. (2024), as displayed in Table 7. We adjust the output instructions to fit our choice response generation method and add all questions as instructions in the ■ battery presentation. For all models we use the default hyperparameter settings.

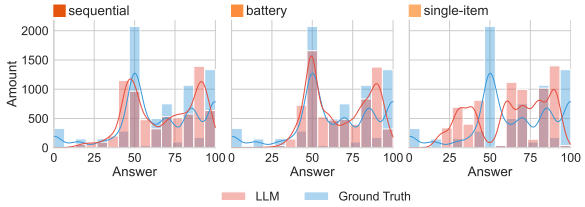

Figure 5: **Answer Distributions.** Predictions and Ground Truth Distributions across the whole population compared for LLama-3.3-70B-Instruct and the question "How do you feel towards Gays and Lesbians?". We can see a clear shift towards the middle for this question, when models are given context of the previous questions and answers, which aligns more closely with human answers.

Attribute	Category
Race	Non-Hispanic White
	Non-Hispanic Black
	Hispanic
Gender	Male
	Female
Ideology	Extremely Liberal
	Liberal
	Slightly Liberal
	Moderate
	Slightly Conservative
	Conservative
	Extremely Conservative

Table 6: **Subpopulations:** We consider these subpopulations for analysis. We have the same subpopulations as the initial study by Bisbee et al. (2024).

B Response Generation OLS Regressions

We obtain the subpopulation-level alignment for each simulation specification and subpopulation, as described in Section 4.3. To identify significant differences in survey response alignment between the response generation methods, we fit the following OLS regression model separately on each dataset (see Table 9): We use the per-subpopulation total variation distance (↓) as the dependent variable and Survey Response Generation Method (reference: Restricted Choice), LLM (reference: Llama 8B), and prompt perturbation (reference: Full Text response options) as independent variables. We use cluster-robust SEs, clustering by seed × decoding strategy, which allows for arbitrary correlation and heteroskedasticity within clusters while assuming independence across clusters. This appropriately

Component	Content
System Prompt	"It is {year}. You are a {age} year-old, {marst}, {raceth} {gender} with {education} making {income} per year, living in the United States. You are {ideo}, {regis} {PID} who {interest} pays attention to what's going on in government and politics."
User Prompt	"Provide responses from this person's perspective. Use only knowledge about politics that they would have. The following questions ask about individuals' feelings toward different groups. Responses should be given on a scale from 0 (meaning cold feelings) to 100 (meaning warm feelings). Ratings between 50 degrees and 100 degrees mean that you feel favorable and warm toward the group. Ratings between 0 degrees and 50 degrees mean that you don't feel favorable toward the group and that you don't care too much for that group. You would rate the group at the 50 degree mark if you don't feel particularly warm or cold toward the group. Output Instructions Questions"
Output Instruction	You only respond in the following JSON format: ```json { "temperature": <temperature> } ```
Output Instruction (■battery)	You only respond in the following JSON format: ```json { "temperature_The Democratic Party?": <temperature_The Democratic Party?>, "temperature_The Republican Party?": <temperature_The Republican Party?>, "temperature_Democrats?": <temperature_Democrats?>, "temperature_Republicans?": <temperature_Republicans?>, "temperature_Black Americans?": <temperature_Black Americans?>, "temperature_White Americans?": <temperature_White Americans?>, "temperature_Hispanic Americans?": <temperature_Hispanic Americans?>, "temperature_Asian Americans?": <temperature_Asian Americans?>, "temperature_Muslims?": <temperature_Muslims?>, "temperature_Christians?": <temperature_Christians?>, "temperature_Immigrants?": <temperature_Immigrants?>, "temperature_Gays and Lesbians?": <temperature_Gays and Lesbians?>, "temperature_Jews?": <temperature_Jews?>, "temperature_Liberals?": <temperature_Liberals?>, "temperature_Conservatives?": <temperature_Conservatives?>, "temperature_Women?": <temperature_Women?> } ```
Question	How do you feel towards the Republican Party?

Table 7: **Prompt.** We use the same prompts for ■sequential and ■single-item and a slightly modified output instruction for the ■battery presentation. For ■battery presentation we ask all questions separated by new lines.

reflects the repeated-measures structure of our evaluation. We do not include interaction terms into the OLS model to mitigate multicollinearity—all VIF values are < 3. We apply Benjamini–Hochberg correction across all reported coefficients in all datasets. Key coefficients for the Verbalized Distribution Method, as well as OLMo 32B and Qwen 32B remain significant even under Bonferroni correction, although Bonferroni is known to be overly conservative in regression settings with correlated predictors.

		(1)	(2)
		MAE (OLS)	**WD Score (WLS)**
Questionnaire Presentation	■ sequential	0.362**	-0.546*
	■ battery	-0.199**	-1.166**
Model	Llama 3.1 8B	2.999**	-0.617*
	Llama 3.2 1B	17.822**	12.796**
	Llama 3.2 3B	6.450**	1.152**
	Phi-4 Mini	3.366**	-0.163
	Qwen3 30B	0.420**	0.488
	Qwen3 4B	2.187**	0.454
	Gemma 3 12B	1.245**	1.713**
	Gemma 3 27B	0.142**	0.448
	Gemma 3 4B	2.074**	1.662**
Interactions (Presentation × Model)	■ sequential × Llama 3.1 8B	-0.216**	0.060
	■ battery × Llama 3.1 8B	0.213**	-0.044
	■ sequential × Llama 3.2 1B	-5.162**	-8.311**
	■ battery × Llama 3.2 1B	-5.273**	-8.203**
	■ sequential × Llama 3.2 3B	-0.476**	-2.195**
	■ battery × Llama 3.2 3B	-1.136**	-1.211**
	■ sequential × Phi-4 Mini	-0.619**	-2.322**
	■ battery × Phi-4 Mini	-1.318**	-1.183**
	■ sequential × Qwen3 30B	-0.973**	-0.784*
	■ battery × Qwen3 30B	-0.418**	-0.372
	■ sequential × Qwen3 4B	-1.125**	-0.885*
	■ battery × Qwen3 4B	0.490**	1.574**
	■ sequential × Gemma 3 12B	-1.218**	-1.366**
	■ battery × Gemma 3 12B	-0.843**	-1.832**
	■ sequential × Gemma 3 27B	-0.779**	-0.625
	■ battery × Gemma 3 27B	0.087	0.250
	■ sequential × Gemma 3 4B	0.652**	0.637
	■ battery × Gemma 3 4B	2.175**	2.394**

Table 8: **Regression Results for MAE and Wasserstein Distance.** (↓) Model (1) uses OLS on Mean Absolute Error. Model (2) uses WLS on Wasserstein Distance, weighted by subpopulation count. Significance levels are based on Benjamini–Hochberg corrected p-values. We can see significant effects for both the questionnaire presentation, but also for the interaction between smaller models and the presentation. Reference categories: *Presentation:* ■ *single-item*, *Model: Llama-3.3-70B-Instruct*. * $p < 0.05$, ** $p < 0.01$

.

		ANES 2016	GLES 2017	GLES 2025	ATP 2021
Intercept		0.374**	0.312**	0.288**	0.503**
Response Generation Method	■ First-Token Probabilities	-0.003	0.147**	0.194**	-0.049*
	■ First-Token Restricted	0.064**	0.220**	0.234**	-0.005
	■ Answer Prefix	-0.002	0.047*	0.085**	-0.082**
	■ Restricted Reasoning	0.017	-0.035*	-0.026	-0.084**
	■ Verbalized Distribution	-0.074**	-0.057**	-0.013	-0.168**
	■ Open-Ended Classif.	0.026	-0.011	-0.027	-0.051**
	■ Open-Ended Distrib.	-0.006	-0.052**	-0.037*	-0.082**
Model	Llama 3B	-0.051*	0.031	0.066**	-0.039*
	Llama 70B	-0.052*	-0.089**	-0.127**	0.007
	OLMo 1B	-0.023	0.109**	0.114**	0.109**
	OLMo 7B	-0.062**	0.070**	0.077**	-0.030
	OLMo 32B	-0.070**	-0.073**	-0.109**	0.016
	Qwen 8B	0.016	0.020	-0.050*	0.075**
	Qwen 8B with Reasoning	-0.012	0.002	-0.010	0.019
	Qwen 32B	-0.076**	-0.108**	-0.161**	-0.036*
	Qwen 32B with Reasoning	-0.056**	-0.067**	-0.081*	-0.106**
Response Option Variants	Full Text, Reversed	0.001	-0.005	0.037	-0.003
	Indexed	0.010	0.003	0.000	0.022*
	Indexed, Reversed	0.035*	0.011	0.026	0.030**

Table 9: **Impact of ■ Response Generation Methods on Subpopulation-Level Alignment (↓).** OLS regression coefficients by dataset with total variation distance (↓) as the dependent variable and Survey Response Generation Method (reference: Restricted Choice), LLM (reference: Llama 8B), and prompt perturbation (reference: Full Text response options) as independent variables. **The Verbalized Distribution Method and larger models lead to significant improvements.** * $p < 0.05$, ** $p < 0.01$ (Benjamini–Hochberg corrected)