Prompt Perturbations Reveal Human-Like Biases in Large Language Model Survey Responses

Jens Rupprecht[1], Georg Ahnert[1], and Markus Strohmaier[1,2,3]
[1]University of Mannheim, Mannheim
[2]GESIS – Leibniz Institute for the Social Sciences, Cologne
[3]Complexity Science Hub, Vienna
[firstname.lastname]@uni-mannheim.de

Abstract

Large Language Models (LLMs) are increasingly used as proxies for human subjects in social science surveys, but their reliability and susceptibility to known human-like response biases, such as *central tendency, opinion floating* and *primacy bias* are poorly understood. This work investigates the response robustness of LLMs in normative survey contexts—we test 18 LLMs on questions taken from the World Values Survey (WVS), applying a comprehensive set of ten perturbations to both question phrasing and answer option structure, resulting in over 334,800 simulated survey interviews. In doing so, we not only reveal LLMs' vulnerabilities to perturbations but also show that almost all tested models exhibit a consistent *recency bias*, disproportionately favoring the last-presented answer option. While larger models are generally more robust, all models remain sensitive to semantic variations like paraphrasing and to combined perturbations. This underscores the critical importance of prompt design and robustness testing when using LLMs to generate synthetic survey data.

1 Introduction

Problem Large Language Models (LLMs) are increasingly being used as proxies for human subjects in social science research, particularly to generate synthetic responses to survey questions (Argyle et al., 2023; Bisbee et al., 2024, inter alia). This application holds promise in augmenting or replacing costly human data collection. Still, the reliability of these synthetic respondents and the extent of overlap with human responses and response biases remain open questions. In particular, research in survey methodology has found that human responses are sensitive to subtle variations in question and answer phrasing that lead to well-known **response biases** (Krosnick, 1991) and it remains unclear whether LLMs, trained on vast amounts of human text, exhibit the same vulnerabilities.

Approach We present a large-scale empirical study, with 334,800 survey interviews in total, investigating the response behavior and robustness of 18 different LLMs to normative questions derived from the World Value Survey (WVS; Haerpfer et al., 2022). By developing and applying a comprehensive set of ten perturbations in both the structure of the answer options (Table 1), such as typos or synonyms, and in the question phrasing (Table 4), such as, e.g., changes to response order or scale structure, we answer the following research questions.

1. Do prompt perturbations negatively affect the **robustness** of LLMs when answering closed-ended, normative survey questions?
2. Do LLMs exhibit **human-like response biases** when answering closed-ended, normative survey questions?

Contribution Next to the perturbation framework, we provide a detailed analysis of LLM response robustness, showing that while some models are more robust than others (e.g. `Llama-3.3-70B-Instruct` and `Gemini-1.5-Pro`), all are susceptible to specific perturbation types. Most notably, we find a consistent *recency bias* across almost all tested models, where the last-presented answer option is disproportionately favored by up to 20 times, and even the largest models remain sensitive to changes in question phrasing.

These findings underscore the importance of careful prompt and Q&A design when using LLMs for synthetic survey responses. The perturbation framework serves as a useful baseline for evaluating robustness in survey contexts, and we make the full Q&A dataset publicly available for benchmarking newer or other LLMs. We present an overview of which LLMs exhibited human-like survey response biases (Table 11).

Proceedings of the 7th Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 1–21

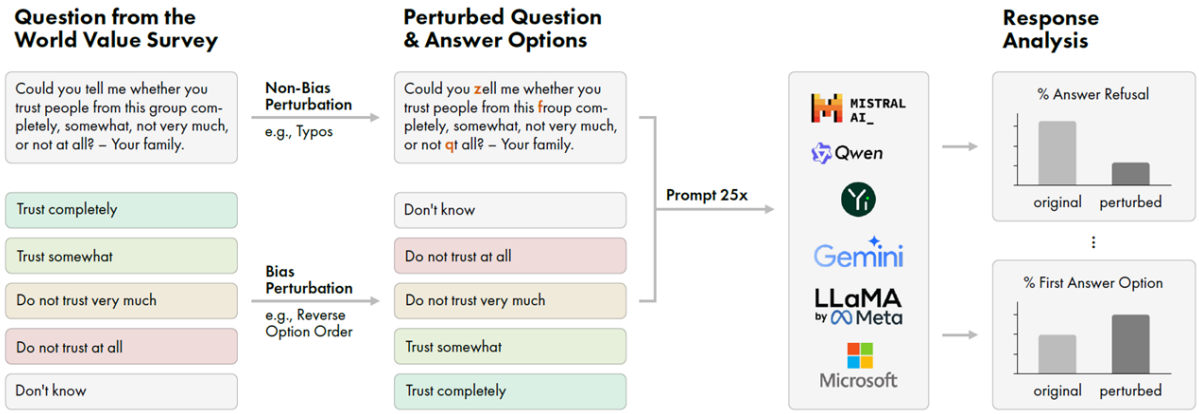


Figure 1: **The Interview Process.** The figure displays an example of a answer option perturbation (a *bias perturbation*, e.g. reversed option order) and an question perturbation (a *non-bias perturbation*, e.g. typos in the question). Each model is prompted 25 times with every perturbation as well as the original Q&A phrasing. All responses are collected, processed and statistically analyzed.

2 Related Work

Our work builds on two main streams of research: (1) survey methodology from the social sciences, which documents human response biases, and (2) recent studies in computer science on the robustness and biases of LLM's synthetic survey response generation.

Human Survey Response Biases Research in the social sciences has shown that how a survey question is asked can be as important as what is asked. Respondents often engage in "satisficing" rather than "optimizing", choosing a satisfactory answer with minimal cognitive effort instead of carefully formulating an optimal one (Krosnick, 1991). This can lead to systematic biases. For example, the order in which the answer options are presented can induce *primacy* (favoring early options in visual surveys) or *recency* (favoring later options in oral surveys) biases (Krosnick and Alwin, 1987). The presence or absence of a middle option or a "don't know" category can trigger a *central tendency bias* or *opinion floating*, respectively (Hollingworth, 1910; Koch and Blohm, 2016). In the first, if a central category is available on the answer scale, humans tend to choose the central category, whereas *opinion floating* indicates that responses are redistributed to central categories if a refusal category is missing (Tjuatja et al., 2024). In addition, *priming* effects, where the preceding context influences subsequent responses, are a well documented phenomenon (Bargh et al., 1996). We draw on these past findings to design perturbations testing whether LLMs exhibit similar human-like response patterns.

LLMs as Survey Respondents Recent studies explored LLMs as substitutes for human survey participants to generate synthetic data. They found that LLMs can replicate average public opinion on political topics, but often with less variance than human samples (Argyle et al., 2023; Bisbee et al., 2024; von der Heyde et al., 2025; Dominguez-Olmedo et al., 2024, inter alia). Others have found that LLM responses can be sensitive to prompting, revealing cultural and demographic biases (Geng et al., 2024). Laverghetta et al. (2022) found that LLMs can produce similar responses to human participants on diagnostic items, for example, in linguistic test. Contrary to this finding, Sühr et al. (2025) found that LLMs' responses to personality tests systematically deviate from human responses. This implies that the results of these tests cannot be interpreted in the same way. However, Huang et al. (2024) identified that LLMs have the potential to represent different personalities with specific prompt instructions. In addition, they found that response patterns of multiple LLMs showed consistency in responses to the Big Five Inventory, indicating a satisfactory level of reliability.

Our work is related to that of Tjuatja et al. (2024), who were among the first to systematically explore human-like response biases in LLMs. They investigated acquiescence, response order, opinion floating, and scale structure effects. Our study extends their work by: (1) using a different, globally diverse survey (the World Values Survey); (2) testing a wider range of LLMs, such as Gemini-2.5-Pro; and (3) incorporating a broader set of perturbations

on both answer and question phrasing, such as *keyboard typos, paraphrasing, synonyms, priming* as well as a combined *interaction* of two perturbations.

LLM Robustness to Perturbations Other researchers have evaluated the general robustness of LLMs to noisy or varied inputs on different tasks. They have shown that even state-of-the-art models can be sensitive to minor changes in the prompt. These perturbations range from the character level, such as typos created by swapping, inserting, or replacing letters (Moradi and Samwald, 2021; Gan et al., 2024), to word- or sentence-level, such as replacing words with synonyms or paraphrasing entire sentences (Qiang et al., 2024). A common finding is that character-level noise can significantly degrade performance, even in large models (Gan et al., 2024). The combination of multiple perturbations can even have a more negative effect (Dong et al., 2023). Although this research has primarily focused on knowledge-based or reasoning tasks, we adapt these perturbation techniques to the context of normative surveys to assess response stability where no single "correct" answer exists.

Evaluation and Prompting Finally, our work is guided by research on the identified ways for evaluating LLMs on multiple-choice tasks. Studies have shown that evaluation results can be highly sensitive to prompt format, e.g. if LLMs face an open- or closed-ended response, and forcing technique. However, forcing a model to choose from a predefined set of options is often necessary to obtain valid responses, as unconstrained responses can differ substantially (Röttger et al., 2024). The returned response labels might differ significantly when a LLM has the option to generate text output before returning the response label due to their auto-regressive nature. Furthermore, relying on the model's first predicted token can misrepresent its full textual output (Wang et al., 2024). Therefore, we include two LLMs with reasoning capabilities to compare their performance to non-reasoning models.

3 Methods

First, we select a subset of 62 questions representing a sample of different thematic categories, each question in the category sharing the same answer options. These normative, value-oriented Q&A pairs are taken from the WVS's core variables (Haerpfer et al., 2022), excluding all sociodemographic variables. Second, we perform the ten perturbations mentioned in Section 3.2 each Q&A pair.

Figure 1 illustrates two exemplary perturbations and the interview process. In total, we perform five perturbations on the answer options as well as five perturbations on the question phrasing of the chosen subset questionnaire. We further include one interaction of two perturbations, one on the answer option and one on the question.

Third, we carry out interviews with the original and each perturbed Q&A pair 25 times with 18 different LLMs. In total, we conducted 334,800 interviews, 18,600 with each model. Last, we compare the distributions of the responded labels for all perturbations response consistency, given the same interview setting, through entropy and calculate the Kullback-Leibler divergence (KL divergence) to compare the baseline response distribution on the original to the perturbed Q&A pairs. *Primacy bias* is further examined by comparing the response frequencies of the first and last answer options in the list, whereas *opinion floating bias* and *central tendency bias* are tested by checking the shift of responses toward or away from the center of the answer option scale.

3.1 Experimental Setup

Survey Data The questions and answer options are sourced from the WVS Wave 7 (2017-2022), a comprehensive cross-national survey on human beliefs and values (Haerpfer et al., 2022). The 259 core WVS Q&A pairs represent 10 distinct thematic categories, including *Trust in People*, *Confidence in Institutions*, *Moral Justifiability*, and *Perception of Democracy*, ensuring a diverse range of topics and answer scale formats (e.g., 3-point to 10-point scales). We used stratified sampling to select six to seven Q&A pairs per thematic category, resulting in a total of 62 Q&A pairs.

Models To ensure that our findings are not specific to a single model architecture or developer, we selected 18 instruction-tuned LLMs, varying in size, developer, and origin, including two with reasoning capabilities. This selection aims to establish external validity for our results and includes proprietary and open-source LLMs. The following models were interviewed:

- **Llama** (in the following tables abbreviated as L3.3-70B (Meta, 2025b), L3.1-8B (Grattafiori

et al., 2024), L3.2-3B, L3.2-1B (Meta, 2025a) represent their respective Instruct versions),

- **Qwen** (Q2.5-7B for Qwen-2.5-7B-Instruct; Q3-0.6B through Q3-32B for Qwen3 versions; Q3-30BT for Qwen3-30B-A3B-Thinking-2507, (Yang et al., 2025b,a)),
- **Gemini** (G1.5P for 1.5-Pro, G2.5P for 2.5-Pro, G2.5F for 2.5-Flash, (Georgiev et al., 2024)), and
- **Others** (M7B for Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), P3.5M for Phi-3.5-mini (Abdin et al., 2024), Y1.5-6B for Yi-1.5-6B-Chat, (Young et al., 2024)).

We listed the specific model IDs in Appendix Table 3.

3.2 Perturbation Design

We designed two categories of perturbations to test model robustness: (1) bias-inducing alterations to the answer options, based on survey methodology research that are known to induce biased responses in humans (Tjuatja et al., 2024), and (2) non-bias alterations to the question phrasing, mimicking common textual variations and errors. Table 1 and Table 4 provide examples as well as references for all perturbations. For each of the 62 Q&A pairs, we created the following ten perturbed versions.

Bias Perturbations These five perturbations manipulate the answer choices provided to test for known survey response biases identified in human subjects and presented in Section 2. Therefore, we call them *bias perturbations* (Tjuatja et al., 2024, p.3).

- **(1) 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').
- **(2) Missing Refusal Option:** The "Don't know" or refusal option is removed from the list of choices.
- **(3) Odd/Even Scale Transformation:** For scales with an even number of options, we use `Gemini-1.5-flash` to generate a semantically appropriate middle category, transforming it into an odd-numbered scale (e.g., adding 'Neutral'). Conversely, for odd-numbered scales, we remove the middle category to create an even-numbered scale and adjust the integer label.
- **(4) Priming Suffix:** A sentence designed to elicit a response is appended to the prompt after the answer options: *'This is very important to my research! You better do not refuse the answer.'*

Non-Bias Perturbations These five perturbations modify the question text to assess robustness to stylistic variations and typos. Typically, humans are unaffected by such subtle changes in the question phrasing and are still able to understand the question's meaning (Tjuatja et al., 2024, p.3). Therefore, we call them *non-bias perturbations*.

- **Typographical Errors:** We introduce three types of typos: **(5) Key Typo** (replacing a character with a random one), **(6) Letter Swap** (swapping two letters in a random word), and **(7) Keyboard Typo** (replacing a character with an adjacent one on a QWERTY keyboard).
- **Semantic Variations:** We use `Gemini-1.5-flash` to create two semantic variations while preserving the original meaning: first, by **(8) Synonym Replacement** where five words in the original question are replaced with synonyms. Second, through **(9) Paraphrasing** the entire question is rephrased.

We manually validated all LLM-generated perturbations (paraphrases, synonyms, odd-scale options) on our 62-question subset to correct errors and ensure their semantic integrity.

Last, we introduce an **(10) Interaction Effect** to study the impact of not only one, but two perturbations. Thus, we created one additional condition that pairs a paraphrased question with reversed-order answer options.

3.3 Interview Procedure and Data Collection

Prompting To ensure internal validity, we used a single, consistent prompt structure for all interviews. The prompt was designed to be clear and direct, instructing the model to select a single option and return only its label.

Answer the following question: {question}
This is a list of possible answer options:
{answer_options}
You must pick one of the answer options. Only answer with the label.

Taking into account the findings of Röttger et al. (2024) we use *forced choice* prompts that provided the most valid responses in their experiments,

Type	Perturbation	Question	Answer Options	Bias and Reference
Original	**Original**	For each of the following aspects, indicate how important it is in your life. Would you say it is very important, rather important, not very important or not important at all? Family	['1=Very important ', '2=Rather important ', '3=Not very important ', '4=Not important at all', "-1=Don't know"]	(Haerpfer et al., 2022)
Bias Perturbations	**(1) Reversed Response Order**	For each of the following aspects, indicate how important it is in your life. Would you say it is very important, rather important, not very important or not important at all? Family	["-1=Don't know", '4=Not important at all', '3=Not very important ', '2=Rather important ', '1=Very important ']	**Primacy Bias** (Tjuatja et al., 2024; Krosnick and Alwin, 1987; Kampen, 2007; O'Halloran et al., 2014)
	(2) Missing Refusal Option		['1=Very important ', '2=Rather important ', '3=Not very important ', '4=Not important at all']	**Opinion Floating Bias** (Schuman and Presser, 2000; Tjuatja et al., 2024)
	(3) Odd/Even Scale Transformation		['1=Very important ', '2=Rather important', '3=Neutral', '4=Not very important ', '5=Not important at all', '-1=Don't know']	**Central Tendency Bias** (Hollingworth, 1910; Cronbach, 1946; Aston et al., 2021; Crosetto et al., 2020)
	(4) Priming Suffix		[1=Very important , 2=Rather important , 3=Not very important , 4=Not important at all, -1=Don't know]This is very important to my research! You better do not refuse the answer.	**Priming Effect** (Bargh et al., 1996; Higgins, 1996; Weingarten et al., 2016; Li et al., 2023)

Table 1: **An exemplary perturbation scheme showing the original question, the bias perturbations and the the hypothesized human-like response bias.** The Q&A example is taken from the thematic category "Importance of Life Aspects" from the WVS wave 7. Bias perturbations have a constant question phrasing and varying answer options. The remaining *Non-Bias Perturbations* are listed in Table 4.

which means that the model correctly chooses one available option from the given answer options. In an unforced prompt setting, LLMs did not follow the instructions and only rarely returned valid response labels, making a statistical evaluation of the results infeasible. Preliminary tests revealed that especially smaller LLMs often fail to perfectly follow the instruction to "answer only with the label", as they return conversational filler or explanations alongside their choice.

Data Collection Each of the 18 models was presented with 12 experimental conditions (1 original + ten perturbations, where for the perturbation *Odd/Even Scale Transformation* one run was done for the odd and even scale scenario) for each of the 62 selected WVS questions. To obtain a stable distribution of responses and enable statistical analysis, we repeated each unique model-Q&A-perturbation combination 25 times. This resulted in a total of $18 \times 62 \times 12 \times 25 = 334,800$ interviews.

Response Extraction and Validation To ensure accurate data for analysis, we developed a robust extraction pipeline. We compared two main approaches. First, `Gemini-1.5-Pro`, `Llama-3.1-8B`, and `Qwen2.5-7B` were prompted, and a regular expression was designed to extract the answer labels. Based on multiple conditions, e.g. if the given answer label is part of the original answer options or that only one response is provided, the methods should highlight which technique is the most promising in extracting valid responses and handling possible edge cases of model responses.

We manually labeled these extraction methods on a random sample of responses for validation. The LLM-based methods achieved accuracies between 77% and 97.5%, with the largest model `Gemini-1.5-Pro` performing best. However, our refined regular expression achieved the overall best extraction success on the validation set as it correctly extracted all responses in our validation set. Consequently, we used this regular expression to process all remaining 316,200 model responses.

4 Results

This section presents the results of our experiments, focusing on two key research questions: (1) LLMs' general robustness to various input perturbations, and (2) their susceptibility to human-like survey response biases revealed by interviews with bias prompt perturbations. In addition, we also investigate the models' general adherence to interview instructions.

4.1 Robustness to Question and Answer Perturbations (RQ1)

We distinguish between response robustness (the tendency to maintain a similar answer distribution under perturbation, measured by KL divergence) and response consistency (the tendency of a model to give the same answer to the same prompt, measured by entropy). A KL divergence of zero indicates a perfect match and thus full robustness against the input perturbation, whereas a high entropy value indicates very inconsistent response behavior.

Effect of Model Size on Robustness First, when assessing robustness to perturbations, we found a clear relationship with model size: *larger models tend to be more robust*. Table 2 and 5 show the percentage of questions for which the models produced a perfectly identical response dis-

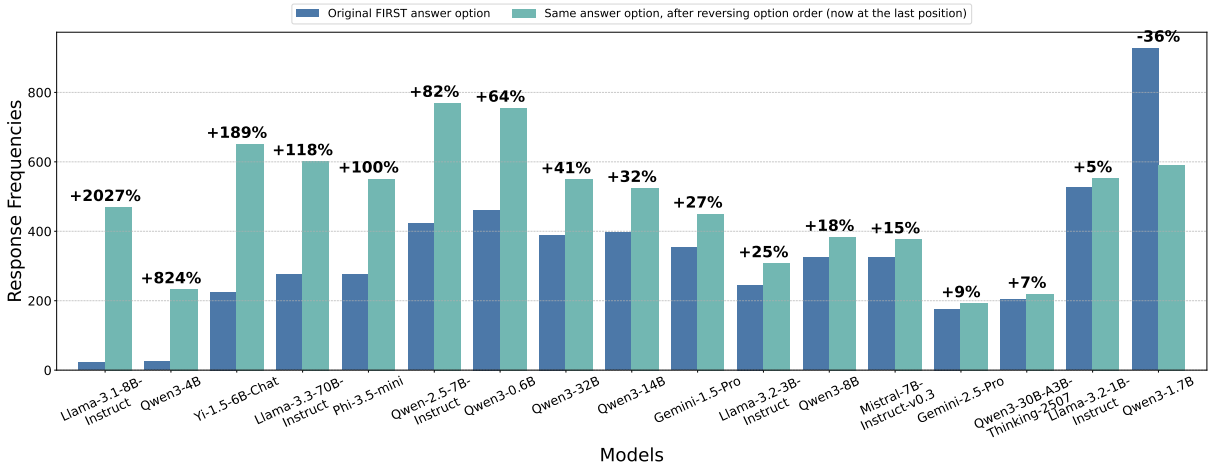


Figure 2: **Evidence of recency bias across all models.** The bars show the frequency of choosing the same answer option (e.g., "Very important") when it is presented first vs. last. Almost all models are more likely to select an option when it appears at the end of the list.

tribution (KL divergence = 0) despite perturbations. `Llama-3.3-70B` and `Gemini-1.5-Pro` were the most robust, often replicating their original answers in over 50% of cases. The smaller `Llama` models were the least robust, with `Llama-3.2-1B` perfectly replicating its answers in fewer than 5% of cases on average. This suggests that scale is a key factor in achieving stable response behavior in synthetic response generation. The reason why the responses for `Gemini-2.5-Pro` and `Flash` are not robust is most likely due to the fact that—despite the same experimental setup—the new Gemini series refuses value-oriented questions much more often than its previous series (see Table 7).

Second, we found that model size is in an inverse relationship with response consistency; smaller models exhibited higher entropy and standard deviation when asked the same question multiple times, indicating more random response behavior (Table 9).

Effect of Perturbation Type on Robustness Further, Tables 2 and 5 highlight the share of fully robust responses (KL divergence = 0) across all questions by perturbation and LLM. It shows that some perturbations had a greater impact on robustness across all models.

- **Combined Perturbations:** The interaction of two perturbations (paraphrased question + reversed answers) has the most bewildering effect on the responses, causing the lowest robustness scores for all models except `Phi-3.5-mini`.
- **Semantic vs. Lexical Changes:** Paraphrasing the question reduced robustness more than replacing individual words with synonyms in most LLMs. These findings are consistent with Moradi and Samwald (2021) who found that models trained on larger corpora are more robust when words are replaced by their synonyms.
- **Typographical Errors:** Randomly replacing characters (*Key Typo*) or using adjacent keys (*Keyboard Typo*) was more robustness-harming than simply swapping two letters within a word (*Letter Swap*). We assume that the training text corpora potentially contain more words with accidental letter swaps than random typos, and therefore LLMs might be more resilient against these perturbations.
- **Answer Option Changes:** Reversing the answer scale or changing it from odd to even (or vice versa) had a more negative impact on the robustness of responses than removing the refusal option or adding emotional priming.

Effect of Answer Option Scale Length We also observed that robustness is affected by the complexity of the task. For nearly all models, the share of fully robust responses decreased as the length of the answer scale, i.e. answer options, increased. For example, models were less likely to replicate their exact response distributions on a 10-point scale compared to a 4-point scale, indicating that a larger decision space can make LLMs more susceptible to perturbations. Figures 4 and 5 suggest that for most LLMs, except Gemini-1.5-pro, the size of the answer option scale has an impact on response robustness comparing the share of fully robust responses on e.g. the 4- and 10-point scale. This suggests

Model	(1) Reversed Answer	(2) Missing Refusal	(3) Even Scale	(4) Priming Suffix
Llama Family				
L3.3-70B	0.50	0.73	**0.60**	**0.82**
L3.1-8B	0.08	0.39	0.27	0.35
L3.2-3B	0.10	0.11	0.16	0.10
L3.2-1B	0.00	0.08	0.03	0.11
Qwen Family				
Q3-32B	0.27	0.35	0.19	0.31
Q3-30BT	0.32	0.23	0.15	0.23
Q3-14B	0.53	0.60	0.48	0.53
Q3-8B	0.34	0.53	0.26	0.39
Q2.5-7B	0.32	0.48	0.45	0.50
Q3-4B	0.24	0.50	0.35	0.31
Q3-1.7B	0.34	0.66	0.47	0.52
Q3-0.6B	0.02	0.03	0.02	0.00
Gemini Family				
G1.5P	**0.69**	0.76	0.55	0.74
G2.5P	0.32	0.11	0.26	0.00
G2.5F	0.15	0.16	0.16	0.00
Others				
M7B	0.68	**0.81**	0.53	0.74
P3.5M	0.53	**0.81**	0.45	0.79
Y1.5-6B	0.47	0.68	0.55	0.52

Table 2: **Share of Fully Robust Responses by Perturbation Type and Model** (↑). The models are grouped by model family.

that the larger the answer scale, the less likely models can reproduce the responses they gave in the original Q&A phrasing, under perturbed settings.

4.2 Evidence of Human-like Survey Biases (RQ2)

With many of the perturbations, we are able to go beyond LLMs robustness and consistency and also analyze whether LLMs exhibit human-like survey response biases. We find evidence for some human-like biases.

Recency Bias Contrary to the initial hypothesized primacy bias, we found *indications of a recency bias in 17 of the 18 models tested.* When we reversed the order of the answer scale, the probability to choose the first option plummeted, while the probability to choose the last option (which is the semantically identical first option in the original Q&A) increased strongly, *ceteris paribus*. As shown in Figure 2, 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. This indicates that LLMs, similar to human respondents in oral surveys, might overemphasize the final options they process. However, LLMs with activated reasoning capabilities, such as `Gemini-2.5-Pro` and `Qwen3-30B.A3B-Thinking-2507`, mitigate this bias. The same answer option placed at the scale end is chosen just 0.07-0.09 times more often after reasoning.

Opinion Floating and Central Tendency The effects of removing the refusal option (*opinion floating*) or providing an explicit middle category (*central tendency*) were highly model-dependent, often correlated with model size (Tables 10a and 10b). For *opinion floating*, larger models like `Llama-70B`, `Gemini-1.5-Pro`, but also `Phi-3.5` were largely robust, showing minimal shifts in their response distributions. Smaller models, particularly `Qwen` and `Llama-8B`, showed a weak tendency to shift responses toward the scale's center when the refusal option was absent. Here, we expect that models redistribute their original non-responses to the center of the answer scale to maintain their indecisiveness, which is also known as opinion floating bias in humans.

Similarly, for *central tendency*, larger models (`Llama-70B`, `Gemini-1.5-Pro`, `Gemini-2.5-Flash`, `Mistral`) consistently shifted their mean response closer to the center across all scale types when an explicit middle option was provided compared to even answer option scales. However, smaller models, such as `Qwen3-0.6B`, `-1.7B`, `-4B` or `Phi-3.5-mini`, showed inconsistent effects or were completely unaffected. Further, we see an almost consistent response shift to the center across all except three models for medium-sized scales (four vs. five point Likert scales).

Binomial tests underlined that the middle option was selected significantly more often than expected under a uniform distribution, especially on larger scales (cf. Table 6). LLMs tend to choose the middle category significantly more often when the scale size increases from three to five and to 11 point Likert scales. All, except the two smallest Qwen3 models and `Llama-3.2-1B`, choose the middle category significantly more often than any other option when facing a eleven answer options.

Emotional Priming The impact of adding an emotional priming statement ("This is very important to my research!") was also model-dependent.

For larger models (Llama-70B, Gemini-1.5-Pro, Mistral), it either had no effect or slightly decreased the rate of refusal responses, suggesting they correctly interpreted the intent of the priming statement. Conversely, for the two Chinese models, Qwen2.5-7B and Yi-1.5-6B, the priming text even *increased* the share of refusal responses across most topics. No clear relationship can be drawn from these findings due to inconsistent behavior across models.

4.3 Interview Adherence and Refusal Rates

Overall, the models demonstrated high adherence to the prompt instructions, with an average of 96% of interviews yielding an extractable and valid answer that was part of the given answer options. However, performance varied significantly across models. Larger models such as Llama-3.3-70B and Gemini-1.5-Pro, but also Phi-3.5-mini and Mistral are very reliable response generators and followed the instructions well while returning little to no incorrect or no answer label. In contrast, other models, particularly smaller Llama models like Llama-3.2-3B (83.6%), Qwen2.5-7B, but also the two reasoning models, were more likely to produce invalid responses that did not follow instructions.

We combined invalid responses with explicit refusals (i.e., choosing the *Don't know* option) to measure overall non-response rates, as shown in Table 7. Llama-3.3-70B, Phi-3.5, and Mistral-7B consistently provided on-scale answers, with non-response rates typically below 10%. Conversely, Qwen2.5-7B and Llama-3.1-8B exhibited high non-response rates, often exceeding 30%.

In particular, we observed topic-specific sensitivity. For questions regarding the *Perception of Elections*, Qwen2.5-7B failed to provide a valid, on-scale answer in 91.3% of cases, even for the original, unperturbed questions. This might suggest the presence of strong content-based guardrails or restrictions in certain models (cf. Figure 8).

5 Discussion and Conclusion

Our experiments revealed that LLMs response robustness is negatively influenced by prompt perturbations when answering closed-ended survey questions (*RQ1*). the variety of perturbation allows us to gain insights into the robustness of LLMs as some models are more sensitive and some perturbations are more robustness-harming than others. For instance, swapping letters within a word has less negative impact than introducing random or keyboard-adjacent characters. This might be explained by the fact that letter swaps are more likely when typing and therefore might potentially take a greater part of training data (Dhakal et al., 2018). This possibly makes the LLM more resilient to this perturbation compared to exchanging characters with random others. Combining two types of perturbations has the strongest negative impact on robustness, whereas synonyms tend to be less confusing than paraphrasing.

Further, we found that perturbations can be an insightful approach to identify human-like survey response biases (*RQ2*). For example, the same answer option is more likely chosen if it is the last mentioned option than if it were the first answer option, holding all other specifications and phrasings constant. This consistent change in the response distribution to the last answer option suggests a *recency bias* rather than a primacy bias.

Although this is not valid across all inspected models, binomial tests reveal that most models choose the middle category more likely than the other categories. Thus, a *central tendency bias* could be identified for specific models across all scale types, whereas none of the LLMs consistently mirrors a *opinion floating bias*.

The findings emphasize the importance of the positioning of answer options when generating synthetic data. In addition, our results highlight the strong sensitivity of LLMs to simple prompt perturbation. Therefore, we strongly recommend researchers to consider prompt robustness checks when deploying closed-ended questions to LLMs. This is because (i) models show very different response behavior and robustness depending on their size, release date (e.g. Gemini-1.5-Pro vs. Gemini-2.5-Pro) and perturbation type, and (ii) LLM response biases are sometimes but not necessarily aligned with biases identified in humans.

Recommendations Based on our findings, we recommend researchers to:

- Use larger, non-reasoning, LLMs for overall better consistency and robustness in generating synthetic survey responses (cf. Tables 2 and 5)
- Reasoning seems to mitigate the *recency bias* identified in non-reasoning LLMs. However, it leads to more non-responses and refusals.
- Use smaller answer option scales for better reproducibility of results (cf. Figure 4).

- Reflect on the meaningfulness of adding a middle category. Including a middle category might steer some LLM responses to the center (cf. Table 10a).
- Reflect the meaningfulness of adding a refusal category. Adding a refusal category might highlight LLM guardrails or restrictions in some thematic areas, as the model can refuse to answer while still following the instructions as it returns a valid response label (cf. Figure 8).
- Use *forced-choice* prompts to generate high turnouts while also considering open-ended evaluation if sensible.

Limitations

This study investigates the robustness of LLM-generated survey responses when facing diverse prompt perturbations, but several methodological and conceptual limitations must be noted. The use of a multiple-choice format, originally designed for human respondents, imposes an artificial constraint on LLMs that typically work in open-ended contexts. As a result, the findings may not generalize to more naturalistic human-LLM interactions.

Although we constrained and validated the data augmentation process, relying on a LLM (`Gemini-1.5-flash`) for generating paraphrases risks semantic drift, as also noted by Qiang et al. (2024). More granular validation—e.g., with multiple human raters—could improve semantic reliability. In addition, perturbations were applied at a fixed intensity, limiting insight into how different degrees of linguistic noise affect model behavior.

Further constraints arise from our prompting and generation setup. The validation set for answer extraction was relatively small compared to the full dataset, so some extraction errors may remain. We also did not apply prompting strategies like persona prompting, shown to improve contextual consistency (Bisbee et al., 2024; Cho et al., 2024), nor used techniques such as *Chain of Thought* prompting. This could promote more deliberative responses instead of only the latent baseline behavior of LLMs. For example, one could use empirically grounded, survey-derived persona collections to gain more perspectivist synthetic survey responses (Rupprecht et al., 2026). Note that practitioners using demographic personas may observe different bias patterns, and that extending the perturbation framework to persona-conditioned settings is an important future direction. Moreover, our experiments focused exclusively on fine-tuned models, leaving open the question of how base models would behave under similar conditions. Additionally, a constant temperature setting restricted our ability to examine variability and creativity in the output.

Finally, reproducibility is another significant challenge. Closed-source LLMs can change without notice, altering response distributions over time and complicating replication efforts, as highlighted by Bisbee et al. (2024). This may have affected our `Gemini` results. Related work also shows that LLMs often offer contradictory answers to semantically equivalent questions when the format shifts from multiple choice, close-ended to an open-ended form (Röttger et al., 2024). Such response instability suggests that observed "attitudes" may be artifacts of prompt design rather than indicators of stable model beliefs or traits.

Ethical Considerations

Generating synthetic survey responses might be relevant in various domains and applied to different use cases, e.g. for pre-testing surveys. However, generating synthetic responses instead of the surveying a real, might result in over-reliance on synthetic responses. This can become risky when there is no ground truth data of the real target population available as the alignment of the artificial responses cannot be evaluated. Frequent reliance on artificial responses may normalize their use where human perspectives are irreplaceable (e.g. in policymaking or clinical trials). This risks sidelining real human voices in domains directly impacting human lives.

Researchers should also consider ethical evasion as one possible issue with synthetic survey responses. Synthetic respondents might be viewed as a way to bypass obligatory ethical review processes since no real human participants are involved. This might encourage under-regulated research practices and in the long run weaken ethical safeguards.

Running inference on the 18 LLMs required significant GPU hours, especially including the initial test phase before finalizing the interview pipeline, raising concerns about the environmental impact of experimenting with synthetic survey responses and the access disparities between well-funded and resource-constrained institutions.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, and Harkirat Behl. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *Preprint*, arXiv:2404.14219.

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.

Stacey Aston, James Negen, Marko Nardini, and Ulrik Beierholm. 2021. Central tendency biases must be accounted for to consistently capture Bayesian cue combination in continuous response data. *Behavior Research Methods*, 54(1):508–521.

John A. Bargh, Mark Chen, and Lara Burrows. 1996. Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, 71(2):230–244.

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.

Suhyun Cho, Jaeyun Kim, and Jang Hyun Kim. 2024. LLM-Based Doppelgänger Models: Leveraging Synthetic Data for Human-Like Responses in Survey Simulations. *IEEE Access*, 12:178917–178927.

Lee J. Cronbach. 1946. Response Sets and Test Validity. *Educational and Psychological Measurement*, 6(4):475–494.

Paolo Crosetto, Antonio Filippin, Peter Katuščák, and John Smith. 2020. Central tendency bias in belief elicitation. *Journal of Economic Psychology*, 78:102273.

Vivek Dhakal, Anna Maria Feit, Per Ola Kristensson, and Antti Oulasvirta. 2018. Observations on Typing from 136 Million Keystrokes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 1–12, New York, NY, USA. Association for Computing Machinery.

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.

Guanting Dong, Jinxu Zhao, Tingfeng Hui, Daichi Guo, Wenlong Wang, Boqi Feng, Yueyan Qiu, Zhuoma Gongque, Keqing He, Zechen Wang, and Weiran Xu. 2023. Revisit Input Perturbation Problems for LLMs: A Unified Robustness Evaluation Framework for Noisy Slot Filling Task. In *Natural Language Processing and Chinese Computing*, pages 682–694, Cham. Springer Nature Switzerland.

Esther Gan, Yiran Zhao, Liying Cheng, Yancan Mao, Anirudh Goyal, Kenji Kawaguchi, Min-Yen Kan, and Michael Shieh. 2024. Reasoning Robustness of LLMs to Adversarial Typographical Errors. *Preprint*, arXiv:2411.05345.

Mingmeng Geng, Sihong He, and Roberto Trotta. 2024. Are Large Language Models Chameleons? An Attempt to Simulate Social Surveys. *Preprint*, arXiv:2405.19323.

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

Matthew Gereti, Alejandro Robinson, Sebastian Williams, Christopher Anderson, and Dominic Walker. 2024. Token-Based Prompt Manipulation for Automated Large Language Model Evaluation.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, and Alex Vaughan. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, and Bi Puranen. 2022. World Values Survey Wave 7 (2017-2022) Cross-National Data-Set.

Matthias Hagen, Martin Potthast, Marcel Gohsen, Anja Rathgeber, and Benno Stein. 2017. A Large-Scale Query Spelling Correction Corpus. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1261–1264, Shinjuku Tokyo Japan. ACM.

Edward Tory Higgins. 1996. Knowledge activation: Accessibility, applicability, and salience. In *Social Psychology: Handbook of Basic Principles*, pages 133–168. The Guilford Press, New York, NY, US.

H. L. Hollingworth. 1910. The Central Tendency of Judgment. *The Journal of Philosophy, Psychology and Scientific Methods*, 7(17):461.

Jen-tse Huang, Wenxiang Jiao, Man Ho Lam, Eric John Li, Wenxuan Wang, and Michael Lyu. 2024. On the Reliability of Psychological Scales on Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6152–6173, Miami, Florida, USA. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *Preprint*, arXiv:2310.06825.

Jarl K. Kampen. 2007. The Impact of Survey Methodology and Context on Central Tendency, Nonresponse and Associations of Subjective Indicators of Government Performance. *Quality & Quantity*, 41(6):793–813.

Achim Koch and Michael Blohm. 2016. Nonresponse Bias (GESIS Survey Guidelines)Nonresponse Bias (GESIS Survey Guidelines). Technical report, GESIS - Leibniz Institute for the Social Sciences.

Jon A. Krosnick. 1991. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5(3):213–236.

Jon A. Krosnick and Duane F. Alwin. 1987. An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement. *Public Opinion Quarterly*, 51(2):201.

Antonio Laverghetta, Animesh Nighojkar, Jamshidbek Mirzakhalov, and John Licato. 2022. Predicting Human Psychometric Properties Using Computational Language Models. In *Quantitative Psychology*, pages 151–169, Cham. Springer International Publishing.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2023. Large Language Models Understand and Can be Enhanced by Emotional Stimuli. *Preprint*, arXiv:2307.11760.

Llama Meta. 2025a. Llama 3.2 Model Card. https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/MODEL_-CARD.md.

Llama Meta. 2025b. Llama 3.3 Model Card. https://github.com/meta-llama/llama-models/blob/main/models/llama3_3/MODEL_-CARD.md.

Milad Moradi and Matthias Samwald. 2021. Evaluating the Robustness of Neural Language Models to Input Perturbations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1558–1570, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alissa O'Halloran, S. Sean Hu, Ann Malarcher, Robert McMillen, Nell Valentine, Mary A. Moore, Jennifer J. Reid, Natalie Darling, and Robert B. Gerzoff. 2014. Response order effects in the Youth Tobacco Survey: Results of a split-ballot experiment. *Survey practice*, 7(3):5.

Yao Qiang, Subhrangshu Nandi, Ninareh Mehrabi, Greg Ver Steeg, Anoop Kumar, Anna Rumshisky, and Aram Galstyan. 2024. Prompt Perturbation Consistency Learning for Robust Language Models. *Preprint*, arXiv:2402.15833.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Rose Kirk, Hinrich Schütze, and Dirk Hovy. 2024. Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models. *Preprint*, arXiv:2402.16786.

Jens Rupprecht, Leon Froehling, Claudia Wagner, and Markus Strohmaier. 2026. German General Social Survey Personas: A Survey-Derived Persona Prompt Collection for Population-Aligned LLM Studies. pages 1761–1780, Palma, Mallorca, Spain.

Howard Schuman and Stanley Presser. 2000. *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*, nachdr. edition. Sage Publ, Thousand Oaks, Calif.

Tom Sühr, Florian E. Dorner, Samira Samadi, and Augustin Kelava. 2025. Challenging the Validity of Personality Tests for Large Language Models. In *Proceedings of the 5th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '25, pages 74–81, New York, NY, USA. Association for Computing Machinery.

Lindia Tjuatja, Valerie Chen, Sherry Tongshuang Wu, Ameet Talwalkar, and Graham Neubig. 2024. Do LLMs exhibit human-like response biases? A case study in survey design. *Preprint*, arXiv:2311.04076.

Leah von der Heyde, Anna-Carolina Haensch, and Alexander Wenz. 2025. Vox Populi, Vox AI? Using Language Models to Estimate German Public Opinion. *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. *Preprint*, arXiv:2402.14499.

Evan Weingarten, Qijia Chen, Maxwell McAdams, Jessica Yi, Justin Hepler, and Dolores Albarracín. 2016. From primed concepts to action: A meta-analysis of the behavioral effects of incidentally presented words. *Psychological Bulletin*, 142(5):472–497.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao

Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025a. Qwen3 Technical Report. *Preprint*, arXiv:2505.09388.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, and Haoran Wei. 2025b. Qwen2.5 Technical Report. *Preprint*, arXiv:2412.15115.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, and Jing Chang. 2024. Yi: Open Foundation Models by 01.AI. *Preprint*, arXiv:2403.04652.

Shengyao Zhuang and Guido Zuccon. 2021. Dealing with Typos for BERT-based Passage Retrieval and Ranking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2836–2842, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Reproducibility Materials

A.1 Infrastructure

The experiments were carried out on a high-performance computing cluster and a local server equipped with NVIDIA H100 (80GB) GPUs. The total runtime for one model's 18,600 interviews, e.g. `Llama-3.1-8B-Instruct` including all perturbed and original Q&As, was ca. 35 minutes with approximately 0.11 seconds per interview. To accommodate larger models on available hardware, we applied 8-bit quantization to `Llama-3.1-8B-Instruct` and `Llama-3.3-70B-Instruct`. Smaller models were run without quantization. Experiments with `Gemini` models were conducted on the Google Cloud Service, Vertex AI. The temperature in all models was kept at their default setting. However, varying the temperature could have a relevant impact and can be varied for robustness checks. The code is made available in an anonymous repository for replication at `https://shorturl.at/NJf6h`.

A.2 Models

This table provides an overview of the open-source LLMs used in the experiments including their Model ID on Huggingface.

Short Name	Huggingface Model ID
Llama 1B	`meta-llama/Llama-3.2-1B-Instruct`
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`
Qwen 7B	`Qwen/Qwen2.5-7B-Instruct`
Qwen 0.6B	`Qwen/Qwen3-0.6B`
Qwen 1.7B	`Qwen/Qwen3-1.7B`
Qwen 4B	`Qwen/Qwen3-4B`
Qwen 8B	`Qwen/Qwen3-8B`
Qwen 14B	`Qwen/Qwen3-14B`
Qwen 32B	`Qwen/Qwen3-32B`
Qwen 30B (R)	`Qwen/Qwen3-30B-A3B-Thinking-2507`
Mistral 7B	`mistralai/Mistral-7B-Instruct-v0.3`
Yi 1.5B	`01-ai/Yi-1.5-6B-Chat`
Phi 3.5B	`microsoft/Phi-3.5-mini-instruct`

Table 3: **Language Models.** We evaluate all Survey Response Generation Methods on 18 open-weight LLMs. LLMs with activated reasoning capabilities are denoted with (R).

B Perturbation Scheme Summary

The following tables describe the remaining non-bias perturbations not introduced in Section 3.2.

The "Type" column categorizes the perturbations into two main classes: "Non-bias Perturbation" and

Type	Perturbation	Question	Answer Options	Bias and Reference
Non-bias Perturbation	**(5) Key Typo**	nor eaca jf the following aspecto, indicete how important it ir wn your liae. Would bou say it is very imporcant, rathes importano, not very imporgant ob not impodtant at all? Famizy	['1=Very important ', '2=Rather important ', '3=Not very important ', '4=Not important at all', "-1=Don't know"]	(Dong et al., 2023; Moradi and Samwald, 2021)
	(6) Letter Swap	For each of the following sapects, indicate how important it is in your life. uoWld you yas it is evry important, ratreh important, ton very important or not important ta all? Family		(Hagen et al., 2017; Moradi and Samwald, 2021; Zhuang and Zuccon, 2021)
	(7) Keyboard Typo	For esch of rhe following aspects, indicate how important ut is un your lide. Would you say it ia very imporrant, rather importamt, nit very important ir nor important ay all? Fanily		(Gan et al., 2024; Zhuang and Zuccon, 2021)
	(8) Synonym Replacement	Crucial in life: Family For each of the following aspects, indicate how significant it is in your life. Would you say it is very important, rather important, not very important or not at all important? Family		(Qiang et al., 2024; Gereti et al., 2024)
	(9) Paraphrase	How important is family to you? Please rate its significance in your life on a scale of "very important" to "not important at all".		(Dong et al., 2023; Qiang et al., 2024)
Interaction	**(10) Paraphrase x Reversal**	How important is family to you? Please rate its significance in your life on a scale of "very important" to "not important at all".	["-1=Don't know", '4=Not important at all', '3=Not very important ', '2=Rather important ', '1=Very important ']	(Dong et al., 2023)

Table 4: **An exemplary perturbation scheme showing non-bias and interaction perturbations.** The example is taken from the item set of category "Importance of Life Aspects". In the WVS wave 7 it is question Q1. Non-bias perturbations have variation in the question phrasing with constant answer options, while the interaction perturbation varies both.

"Interaction." The "Perturbation" column specifies the exact modification technique applied, which includes methods such as "Key Typo," "Letter Swap," "Keyboard Typo," "Synonym Replacement," and "Paraphrase." The "Question" column displays the resulting text after each specific perturbation is applied to the original question about the importance of family. The "Answer Options" column lists the response scale provided to the survey participant. Finally, the "Bias and Reference" column provides citations to relevant scientific literature for each perturbation type.

In the first five perturbations, the phrasing of the question is intentionally altered—for instance, by introducing typographical errors (e.g., "Key Typo," "Letter Swap"), substituting words with similar meanings ("Synonym Replacement"), or rephrasing the entire sentence ("Paraphrase"). While the question varies, the "Answer Options" remain constant, consistently ranging from "1-Very important" to "4-Not important at all.". In "(10) Paraphrase x Reversal," the question is paraphrased, and simultaneously, the order of the "Answer Options" is inverted, starting with "4-Not important at all" and ending with "1-Very important."

C Results

This section summarizes the findings discussed in the main part of the work and can serve as a reference to identify other response patterns as the heatmaps contain much information regarding LLMs, perturbation type as well as additional statistical tests on the response distributions.

C.1 Robustness against Non-Bias Perturbations

This heatmap highlights to which extent the LLMs are affected by *non-bias perturbations.* We can see large model-specific differences.

We see that especially the smallest `Llama` models are responding not in a robust way. These results are consistent across the different *non-bias perturbations*. Especially when facing more than one perturbation in the *Interaction* perturbation, where both answer option scale and question phrasing were altered, the robustness drops drastically across all models .

Moreover, we identify perturbations that are less robustness-harming than others. For example, swapping letters within a word does not impact response robustness as much as typos or introducing completely different words, synonyms, or rephrasing the whole sentence.

C.2 Distance Calculation for Central Tendency and Opinion Floating Bias

This section shows how the response shifts to the central category is measured for the bias perturbations *Odd/Even Scale Transformation* and *Priming Suffix*. By calculating the differences in distances to the central point of the answer option scale we try to identify if the average distribution shifts to the central scale point. The actual differences in distance for each answer option scale type and for the two perturbations are visualized in Table 10a and Table 10b.

To better understand how the shift towards the

Model	(5) Key Typos	(6) Letter Swap	(7) Keyboard Typos	(8) Synonyms	(9) Paraphrase	(10) Paraphrase x Reversed
Llama Family						
L3.3-70B	0.52	**0.76**	0.56	0.58	**0.61**	0.44
L3.1-8B	0.32	0.31	0.21	0.31	0.15	0.10
L3.2-3B	0.02	0.11	0.08	0.13	0.05	0.03
L3.2-1B	0.03	0.10	0.00	0.13	0.00	0.00
Qwen Family						
Q3-32B	0.31	0.34	0.31	0.34	0.19	0.23
Q3-30BT	0.19	0.31	0.19	0.31	0.21	0.23
Q3-14B	0.34	0.47	0.37	0.39	0.40	0.39
Q3-8B	0.29	0.35	0.32	0.37	0.37	0.19
Q2.5-7B	0.48	0.63	0.42	0.55	0.44	0.37
Q3-4B	0.39	0.44	0.47	0.50	0.34	0.19
Q3-1.7B	0.50	0.58	0.58	0.53	0.60	0.34
Q3-0.6B	0.03	0.05	0.10	0.13	0.06	0.00
Gemini Family						
G1.5P	**0.68**	0.73	**0.66**	0.58	0.53	0.24
G2.5P	0.35	0.34	0.35	0.31	0.32	0.24
G2.5F	0.21	0.16	0.27	0.10	0.15	0.13
Others						
M7B	0.58	0.65	0.60	**0.71**	0.53	0.45
Y1.5-6B	0.50	0.50	0.45	0.65	0.29	0.29
P3.5M	0.50	0.61	0.47	**0.71**	0.53	**0.48**

Table 5: **Share of Fully Robust Responses by Perturbation Type and Model (↑).** The models are ordered by model family and parameter size.

middle is measured, we present an anecdotal visualization in Figure 3 of the thought behind whether we observe a *central tendency bias* or an *opinion floating bias*.

In addition, Table 6 underlines that a middle category is significantly more often chosen than assumed under a uniform, or random, distribution of responses across the scale.

Thus, a twofold analysis of not only investigating the shift of average responses between response distributions in the perturbation and original setting is important, but also the statistically assessment to make grounded claims.

By conducting a statistical binomial test, we tried to account for that (Table 6).

C.3 Refusal and Invalid Responses

This section should give a broader overview of the refusal rates (LLM chose the "Don't know" answer option) and the invalid responses (e.g. a LLM did not return any valid response).

It is important to have inspect the overall return rates for all the models as these might have implications on the interpretability of the results. For example, when a model exhibits high refusal or invalid response rates, its results might be not very well interpretable as the main analysis only focused on the valid responses.

Model	3-pt Likert Scale	5-pt Likert Scale	11-pt Likert Scale
Llama Family			
L3.3-70B	1.00	1.00	**0.00**
L3.1-8B	1.00	0.07	**0.00**
L3.2-3B	**0.00**	**0.00**	**0.00**
L3.2-1B	0.31	**0.00**	1.00
Qwen Family			
Q3-32B	1.00	0.01	**0.00**
Q3-30BT	1.00	**0.00**	**0.00**
Q3-14B	0.85	0.95	**0.00**
Q3-8B	1.00	0.11	**0.00**
Q2.5-7B	1.00	0.76	**0.00**
Q3-4B	0.59	0.92	**0.00**
Q3-1.7B	**0.00**	1.00	1.00
Q3-0.6B	1.00	1.00	1.00
Gemini Family			
G1.5P	1.00	1.00	**0.00**
G2.5P	1.00	**0.00**	**0.00**
G2.5F	1.00	**0.00**	**0.00**
Others			
M7B	**0.00**	**0.00**	**0.00**
Y1.5-6B	1.00	**0.00**	**0.00**
P3.5M	1.00	1.00	**0.00**

Table 6: **P-Values of a Binomial Test on the Middle Category in an Odd Scale.** The p-values indicate a hypothesis test with a Null-Hypothesis stating that the middle category is not selected significantly more often assumed under a uniform distribution or completely random selection of answer options. However, for larger scale types, the middle category becomes much more relevant as it is significantly chosen more frequently than any other category.

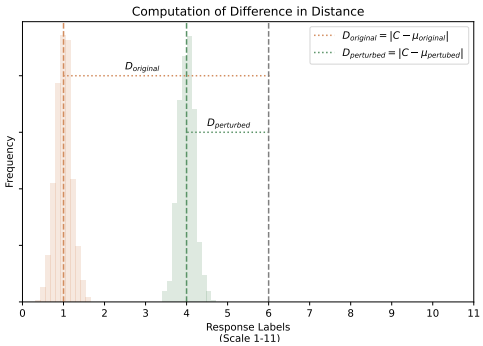


Figure 3: **Exemplary Difference in Distances to Scale Center of Responses to a Perturbed and Original Q&A Pair.** The absolute distance is measured between the scale center and the response mean. Then, $D = D_{perturbed} - D_{original}$. A negative result indicates that the mean response in the perturbed setting is closer to the ideal scale center.

Therefore, this analysis gives insights which results are more reliable than others as for some models there are more valid responses as for others. For example, for `Qwen` we can see large refusal and invalid response rates, generally, but especially in sensitive thematic areas, such as *Perceptions of Elections.*

C.4 Consistency of LLM Survey Responses

This section shows how consistent different LLMs respond to close-ended survey questions when facing the same Q&A pair multiple times. As explained in Section 3, we provide each model with the same Q&A pair in each perturbation stage 25 times and request a response. By running the same setting multiple times we try to identify how consistent LLMs respond generally and whether there are differences when facing the same, but syntactically incorrect (e.g. key typos, etc.), prompts.

Figure 9 highlights that the LLMs consistency in responding is not really affected by specific perturbation types. Thus, "incorrect, flawed prompts" do not increase the response inconsistency of LLMs.

However, there are again large differences between models. We see that especially the smallest `Llama` models and `Qwen` exhibit strong inconsistent responses given the same Q&A pair 25 times, whereas larger LLMs are very consistent. Nonetheless, the `Llama` model family seems to be more inconsistent, generally.

Model	Orig.	Rev. Ans.	Miss. Ref.	Odd Scale	Even Scale	Emo. Prim.	Key Typos	Letter Swap	Keyb. Typos	Syn-onyms	Para-phrase	Para. x Rev.
Llama Family												
L3.3-70B	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	0.01	**0.00**	0.02	0.06	0.03
L3.1-8B	0.19	0.02	**0.00**	0.14	0.23	0.19	0.25	0.23	0.22	0.25	0.20	0.10
L3.2-3B	0.23	0.23	0.15	0.28	0.22	0.22	0.30	0.37	0.27	0.24	0.30	0.25
L3.2-1B	0.15	0.28	0.07	0.09	0.32	0.19	0.11	0.16	0.15	0.16	0.19	0.24
Qwen Family												
Q3-32B	0.02	0.02	0.01	0.05	0.02	0.10	0.10	0.06	0.08	0.04	0.05	0.06
Q3-30BT	0.50	0.47	0.04	0.48	0.56	0.46	0.49	0.52	0.50	0.45	0.57	0.57
Q3-14B	0.10	0.08	**0.00**	0.07	0.11	0.09	0.22	0.12	0.19	0.14	0.21	0.11
Q3-8B	0.01	0.03	**0.00**	**0.00**	0.01	0.05	0.11	0.09	0.12	0.05	0.10	0.08
Q2.5-7B	0.40	0.08	0.06	0.31	0.43	0.42	0.49	0.46	0.50	0.37	0.38	0.10
Q3-4B	0.16	0.01	**0.00**	0.12	0.23	0.18	0.14	0.11	0.12	0.14	0.08	0.06
Q3-1.7B	0.02	0.03	0.02	0.02	0.07	0.04	0.08	0.02	0.07	0.06	**0.00**	0.05
Q3-0.6B	0.14	0.01	0.02	0.11	0.14	0.24	0.08	0.10	0.09	0.14	0.16	0.01
Gemini Family												
G1.5P	0.12	0.11	0.03	0.03	0.11	0.07	0.14	0.08	0.11	0.14	0.28	**0.00**
G2.5P	0.50	0.59	0.12	0.47	0.60	0.70	0.53	0.50	0.51	0.50	0.51	0.57
G2.5F	0.44	0.41	**0.00**	0.44	0.50	0.47	0.48	0.44	0.48	0.40	0.47	0.47
Others												
M7B	0.06	0.08	0.03	0.06	0.08	0.03	0.03	0.03	0.03	0.06	0.06	0.03
Y1.5-6B	0.06	0.05	**0.00**	0.16	0.18	0.11	0.06	0.02	0.10	0.06	0.35	0.05
P3.5M	0.05	**0.00**	**0.00**	**0.00**	0.10	0.08	0.10	0.06	0.10	0.05	0.02	0.02

Table 7: **Overall Share of Unsuccessful and Refusal Interviews across Perturbation Type and LLMs.** (↓) Especially the largest non-reasoning models, such as Llama-3.3-70B, Qwen3-32B, and Gemini-1.5-Pro do not refuse the answers or generate wrong interviews (e.g. wrong labels). Reasoning, however, drastically reduces the adherence to respond to a question with one of the answer option categories.

C.5 Comparison of Robustness against Perturbations by Scale Size

The following plots try to reveal in more detail the extent of robustness drop by perturbation depending on the answer option scale dimension. We identified a drop in robustness as the scale size became larger.

The responses of the smallest LLMs are generally not robust at all. However, when the scale has ten options the responses robustness plummets across all perturbations and not even a single response distribution returned in the original Q&A setting can be generated. This indicated that the smallest `Llama` as well as the chinese LLMs `Qwen` and `Yi` are not able to cope with perturbations, especially when facing a lot of options to choose from.

In all cases, the interaction perturbation with both question and answer option alterations leads to the largest drop in robustness across all models and scale sizes. It is striking, that larger models, especially state-of-the-art models like `Gemini-1.5-Pro`, cannot answer robustly given multiple perturbations.

Researchers should take into account the scale size when generating synthetic responses from close-ended survey Q&A pairs.

Model	Orig.	Rev. Ans.	Miss. Ref.	Odd Scale	Even Scale	Emo. Prim.	Key Typos	Letter Swap	Keyb. Typos	Syn-onyms	Para-phrase	Para. x Rev.
Llama Family												
L3.3-70B	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
L3.1-8B	0.51	**0.00**	**0.00**	0.39	0.54	0.43	0.71	0.50	0.39	0.72	0.25	**0.00**
L3.2-3B	0.39	0.41	**0.00**	0.46	0.38	0.48	0.45	0.40	0.38	0.39	0.23	0.13
L3.2-1B	0.03	0.02	0.07	0.07	0.07	**0.00**	0.03	**0.00**	0.01	0.06	0.13	0.13
Qwen Family												
Q3-32B	0.05	0.10	**0.00**	0.07	0.07	0.20	0.40	0.23	0.28	0.01	0.09	0.02
Q3-30BT	0.97	0.91	0.13	0.93	0.95	0.99	0.88	0.92	0.79	0.49	0.87	0.79
Q3-14B	**0.00**	**0.00**	**0.00**	0.01	**0.00**	0.21	0.12	0.07	0.05	**0.00**	0.17	0.07
Q3-8B	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Q2.5-7B	0.91	0.34	**0.00**	0.69	0.93	0.50	0.83	1.00	0.83	0.83	0.83	0.67
Q3-4B	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	0.13	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	0.03
Q3-1.7B	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	0.01	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Q3-0.6B	0.01	0.01	**0.00**	0.01	0.01	0.11	0.01	0.07	0.01	0.01	0.13	0.01
Gemini Family												
G1.5P	0.37	0.47	**0.00**	0.05	0.34	**0.00**	0.81	0.33	0.33	0.33	0.17	**0.00**
G2.5P	1.00	1.00	0.31	1.00	1.00	1.00	0.98	1.00	1.00	0.90	0.87	0.81
G2.5F	0.86	0.91	0.02	0.81	0.83	1.00	0.93	0.89	0.89	0.64	0.85	0.87
Others												
M7B	0.17	**0.00**	**0.00**	0.17	0.17	**0.00**	0.17	0.17	0.17	**0.00**	0.33	0.17
Y1.5-6B	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	0.17	**0.00**	**0.00**	**0.00**	**0.00**	0.17	**0.00**
P3.5M	0.17	**0.00**	**0.00**	**0.00**	0.17	0.50	0.17	0.17	**0.00**	0.17	0.17	**0.00**

Table 8: **Perception of Elections: Share of Refusal & Unsuccessful Interviews.** (↓) The models are ordered by model family and parameter size.

Model	Orig.	Rev. Ans.	Miss. Ref.	Odd Scale	Even Scale	Emo. Prim.	Key Typos	Letter Swap	Keyb. Typos	Syn-onyms	Para-phrase	Para. x Rev.
Llama Family												
L3.3-70B	0.13	0.04	0.20	**0.00**	0.14	0.26	0.07	0.20	**0.00**	0.18	0.20	0.08
L3.1-8B	0.61	1.08	0.68	0.44	0.53	0.84	0.14	0.69	0.39	0.65	0.34	0.68
L3.2-3B	1.40	1.19	1.19	1.16	1.35	1.18	1.19	1.25	1.49	1.61	1.46	0.95
L3.2-1B	1.94	1.53	1.92	1.38	1.90	1.34	1.35	1.51	1.15	1.70	1.45	1.21
Qwen Family												
Q3-32B	0.73	0.62	0.66	0.17	0.73	1.00	0.79	0.83	1.01	1.12	0.39	0.49
Q3-30BT	0.39	0.42	0.21	0.32	0.40	0.54	0.28	0.46	0.31	0.43	0.56	**0.00**
Q3-14B	0.21	0.30	0.39	0.03	0.31	0.28	0.24	0.17	0.33	0.37	0.10	0.23
Q3-8B	0.41	0.33	0.46	0.26	0.50	0.33	0.26	0.62	0.27	0.80	0.28	0.44
Q2.5-7B	0.38	0.46	0.38	0.35	0.41	0.42	0.13	0.20	0.16	0.14	0.34	0.34
Q3-4B	0.79	0.69	1.29	0.50	0.83	0.31	0.50	0.44	0.19	0.59	0.67	0.54
Q3-1.7B	0.14	0.03	0.14	0.15	0.16	0.41	0.11	0.03	0.02	0.23	**0.00**	**0.00**
Q3-0.6B	0.68	0.74	1.43	0.91	0.58	2.00	0.58	0.42	0.31	0.58	0.55	0.52
Gemini Family												
G1.5P	**0.00**	**0.00**	0.07	**0.00**	0.02	0.05	0.22	0.17	**0.00**	0.39	**0.00**	**0.00**
G2.5P	0.48	0.50	0.57	0.39	0.47	0.69	0.38	0.61	0.51	0.68	0.33	0.50
G2.5F	0.85	0.64	0.85	0.54	0.86	0.83	0.82	0.83	0.63	1.27	0.50	0.46
Others												
M7B	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Y1.5-6B	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
P3.5M	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**

Table 9: **Large model-specific differences in response entropy.** (↓) Little to no perturbation-specific differences. Each scale size subsumes all selected questions. This figure displays the mean entropy across all questions in that scale type for all perturbation and model combinations. Warmer colors indicate a higher average dispersion of the responses across the potential answer options. E.g., if a model answers always with the same label, the entropy is 0.

Model	3-pt Likert Scale	4-pt Likert Scale	5-pt Likert Scale	10-pt Likert Scale
Llama Family				
L3.3-70B	0.35	**-0.04**	0.04	**-0.12**
L3.1-8B		**-0.19**	**-0.15**	**-2.46**
L3.2-3B		**-0.47**	0.04	**-0.34**
L3.2-1B	**-0.04**	0.14	0.01	**-0.15**
Qwen Family				
Q3-32B	0.04	**-0.05**	0.03	**-0.37**
Q3-30BT	**-0.18**	0.81	0.07	**-0.15**
Q3-14B	**-0.04**	**-0.02**	0.03	**-0.71**
Q3-8B	0.01	**-0.11**	0.03	0.27
Q2.5-7B	0.19	**-0.49**	**-0.01**	0.28
Q3-4B	**-0.32**	**-0.12**	**-0.35**	0.24
Q3-1.7B	**-0.03**	**-0.12**	**-0.47**	0.70
Q3-0.6B	0.20	0.08	**-0.06**	1.49
Gemini Family				
G1.5P	0.06	**-0.26**		0.26
G2.5P	0.10	**-0.07**	**-0.07**	0.38
G2.5-F	0.09	**-0.07**	0.12	0.09
Others				
M7B	0.17	0.04		**-0.23**
Y1.5-6B	**-0.17**	**-0.08**	0.17	2.95
P3.5M	**-0.20**	**-0.05**	0.33	**-0.48**

(a) **Models adjust their answer behavior towards the middle when the refusal category is missing.**

Model	3-pt Likert Scale	4-pt Likert Scale	10-pt Likert Scale
Llama Family			
L3.3-70B	**-0.28**	**-0.22**	**-1.11**
L3.1-8B	**-0.50**	**-0.18**	**-0.13**
L3.2-3B	**-0.05**	**-0.28**	0.39
L3.2-1B	0.31	**-0.53**	0.78
Qwen Family			
Q3-32B	0.18	**-0.16**	**-0.43**
Q3-30BT	**-0.36**	0.78	**-1.06**
Q3-14B	**-0.31**	**-0.24**	**-0.38**
Q3-8B	**-0.28**	**-0.42**	**-1.83**
Q2.5-7B	0.42	**-0.23**	**-2.01**
Q3-4B	0.43	**-0.13**	1.43
Q3-1.7B	0.46	0.13	**-3.25**
Q3-0.6B	**-0.14**	**-0.43**	0.40
Gemini Family			
G1.5P	**-0.33**	**-0.47**	**-1.11**
G2.5P	**-0.61**	0.23	**-0.62**
G2.5F	**-0.21**	**-0.40**	**-0.45**
Others			
M7B	**-0.02**	**-0.28**	**-1.08**
Y1.5-6B	**-0.15**	**-0.16**	**-0.26**
P3.5M	0.30	**-0.26**	1.60

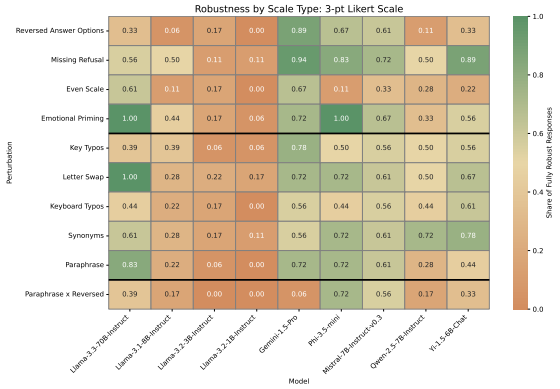
(b) **Models adjust their answer behavior towards the middle when a middle category is existent.**

Table 10: **The values display the difference in mean distance of the perturbed, (a) without refusal category and (b) with middle category to the scale center. Bold values indicate a shift towards the scale center.** For original even scales an artificial middle category is created and vice versa to be able to compare even and odd scales with one another for every question. Thus, in an original 5-pt Likert scale the middle category is removed, whereas in a 4-pt Likert scale a middle category is added. No changes are removed for better readability.

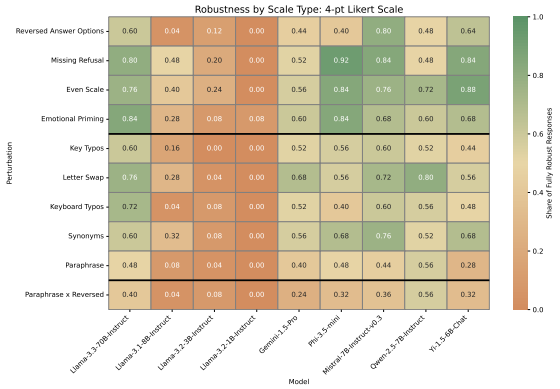
Family	Model	Params	Recency Bias	Opinion Floating	Central Tendency	Emotional Priming
Llama	Llama-3.3-70B-Instruct	70B	✓	×	∼	×
	Llama-3.1-8B-Instruct	8B	✓	∼	∼	×
	Llama-3.2-3B-Instruct	3B	✓	∼	✓	×
	Llama-3.2-1B-Instruct	1B	✓	∼	∼	×
Qwen	Qwen3-32B	32B	✓	∼	∼	×
	Qwen3-30B-A3B (Thinking)	30B	✓	∼	✓	×
	Qwen3-14B	14B	✓	∼	∼	×
	Qwen3-8B	8B	✓	∼	∼	×
	Qwen2.5-7B-Instruct	7B	✓	∼	∼	∼
	Qwen3-4B	4B	✓	∼	∼	×
	Qwen3-1.7B	1.7B	×	∼	∼	×
	Qwen3-0.6B	0.6B	✓	∼	×	×
Gemini	Gemini-1.5-Pro	n/a	✓	×	∼	×
	Gemini-2.5-Pro	n/a	✓	∼	✓	×
	Gemini-2.5-Flash	n/a	✓	∼	✓	×
Others	Mistral-7B-Instruct-v0.3	7B	✓	×	✓	×
	Phi-3.5-mini-Instruct	3.8B	✓	×	✓	×
	Yi-1.5-6B-Chat	6B	✓	∼	∼	∼

Notes: **Recency Bias**: disproportionate selection of the last-presented answer option when response order is reversed (17/18 models affected; increase ranges from $+5\%$ to $+2027\%$). **Opinion Floating**: shift of responses toward the scale centre when a "Don't know" refusal option is removed; most pronounced in smaller models. **Central Tendency**: over-selection of a middle category when one is explicitly provided; significant on larger scales (binomial tests, $p < .05$ for most models on 11-point scales). **Emotional Priming**: change in refusal rate after appending "This is very important to my research! You better do not refuse the answer."; inconsistent across models. Reasoning-capable models show the recency bias at greatly reduced magnitude (≈ 0.07–$0.09\times$ rather than $> 1\times$) but tend to generate more invalid/refusal responses overall.

Table 11: Human-Like Survey Response Biases Identified per LLM. **Legend:** ✓ human-like survey response bias; ∼ partial or inconsistent evidence; × not detected; Model sizes are approximate parameter counts.

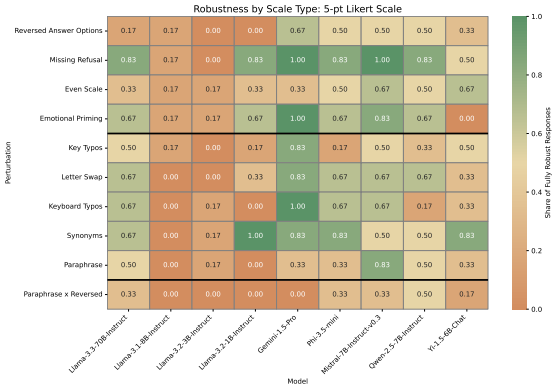


(a) 3-point Likert Scale

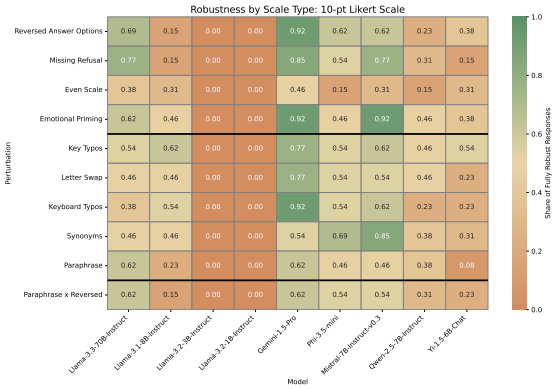


(b) 4-point Likert Scale

Figure 4: **Model-specific differences in fully robust responses on most perturbations on the 3 and 4-point scale.** This figure shows the share of fully robust response distributions given the original response distribution and the responses based on the specific perturbation on the y-axis. Compared to 5 the robustness of responses drops when the scale size becomes larger. The smallest `Llama` models perform very poorly across all scales.



(a) 5-point Likert Scale



(b) 10-point Likert Scale

Figure 5: **Model-specific differences in fully robust responses on most perturbations on the 5- and 10-point scale.** This figure shows the share of fully robust response distributions given the original response distribution and the responses based on the specific perturbation on the y-axis. Compared to 4 the robustness of responses drops when the scale size becomes larger. The smallest `Llama` models perform poorly across all scales.