Prompt Perturbations Reveal Human-Like Biases in LLM 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

Abstract

Large Language Models (LLMs) are increasingly used as proxies for human subjects in social science surveys, but their reliability and susceptibility to known response biases are poorly understood. This paper investigates the response robustness of LLMs in normative survey contexts—we test nine diverse LLMs on questions from the World Values Survey (WVS), applying a comprehensive set of 11 perturbations to both question phrasing and answer option structure, resulting in over 167,000 simulated interviews. In doing so, we not only reveal LLMs' vulnerabilities to perturbations but also reveal that all tested models exhibit a consistent, but variably intense, *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. By applying a set of perturbations, we reveal that LLMs partially align with survey response biases identified in humans. 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 for generating synthetic responses to survey questions (Argyle et al., 2023; Bisbee et al., 2024, inter alia). This application holds promise for augmenting or replacing costly human data collection, but the reliability of these synthetic respondents and to what extent they overlap with human responses remain open questions. In particular, survey methodology research 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 In this paper we present a large-scale empirical study where we investigate the response behavior and robustness of nine different LLMs to normative questions derived from the World Value Survey (WVS; Haerpfer et al., 2022). We systematically apply eleven perturbations, targeting both the question phrasing and the structure of the answer options, as shown in Figure 1. Our research questions are:

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?

We find a consistent *recency bias* across all tested models in different strengths, where the last-presented answer option is disproportionately favored. Larger models generally exhibit more stable response patterns, but even the largest models are sensitive to changes in question phrasing.

Contribution We conduct an experimental evaluation with over 167,000 interviews across nine distinct LLMs, varying in size and origin. We develop and apply a comprehensive set of 11 perturbations, targeting classic survey biases (e.g., response order, scale structure) with common textual variations (e.g., typos, paraphrasing). This makes it a useful baseline to test newly developed LLMs on the same task to put the results into perspective. We provide a detailed analysis of LLM response stability, showing that while some models are more robust than others, all are susceptible to specific types of perturbations. This work underscores the importance of careful prompt and Q&A design when using LLMs as a resource for synthetic survey responses, and provides a framework for evaluating their robustness in survey contexts.[1]

[1]The material can be accessed at https://github.com/dess-mannheim/Survey-Response-Biases-in-LLMs.git.

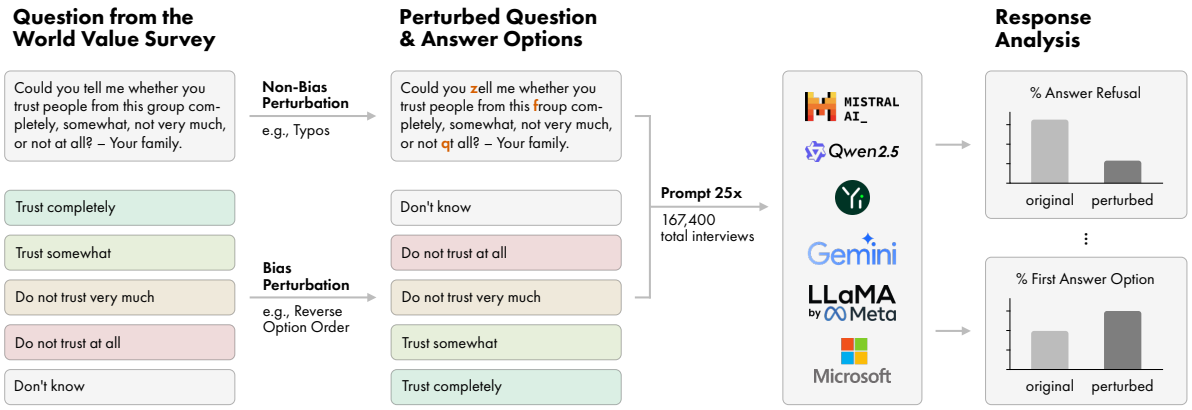


Figure 1: **The Interview Process.** The figure shows an example of a bias perturbation (e.g. reversed option order) and a non-bias perturbation (e.g. typos in the question). Each model is interviewed 25 times with each combination.

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, inter alia). Others have found that LLM responses can be sensitive to prompting, revealing cultural and demographic biases (Geng et al., 2024).

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 of varying sizes and developers; 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 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 answers 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).

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. On each Q&A pair we perform all eleven perturbations.

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.

Interviews with the original and each perturbed Q&A pair are carried out 25 times with nine different LLMs. In total, we conducted 167,400 interviews, 18,600 with each model. We compare the distributions of the responded labels for all perturbations through entropy and KL-divergence to the baseline response distribution to the original Q&A pairs. *Primacy bias* is further examined by comparing the response frequencies of the first and last answer option 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). From the 259 core variables concerning values and norms, we selected a subset of 62 questions. This subset represents 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).

Models To ensure our findings are not specific to a single model architecture or developer, we selected nine instruction-tuned LLMs, varying in size, developer, and origin. This selection aims to establish a degree of external validity for our results and includes proprietary and open-source LLMs. The following models were interviewed: `Gemini-1.5-Pro`, `Llama-3.3-70B-Instruct`, `Llama-3.1-8B-Instruct`, `Llama-3.2-3B-Instruct`, `Llama-3.2-1B-Instruct`, `Mistral-7B-Instruct-v0.3`, `Phi-3.5-mini-instruct`, `Qwen2.5-7B-Instruct` and `Yi-1.5-6B-Chat`.

Infrastructure Experiments were carried out on a high-performance computing cluster (bwUniCluster 2.0) and a local server equipped with NVIDIA H100 (80GB) GPUs. The total runtime on bwUniCluster 2.0 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. `Gemini-1.5-Pro` was accessed via its official API.

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 and known to induce biased responses in humans, and (2) non-bias alterations to the question phrasing, mimicking common textual variations and errors. Appendix Table 1 provides examples

as well as references for all the perturbations. For each of the 62 Q&A pairs, we created the following eleven perturbed versions.

Bias Perturbations These perturbations manipulate the provided answer choices to test for known survey response biases identified in human subjects.

- **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:** 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, 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.
- **Emotional Priming:** 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 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.

- **Typographical Errors:** We introduce three types of typos: *Letter Swap* (swapping two letters in a random word), *Key Typo* (replacing a character with a random one), and *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 **Synonym Replacement** where five words in the original question are replaced with synonyms. Second, through **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.

Interaction Effect To study combined effects, 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 a combination of their *forced choice* prompts that rendered the most valid responses, which means that the model is correctly choosing one available option from the given answer options. In an unforced prompt setting, the model did not follow the instructions and only rarely returned valid response labels, making a statistical evaluation of the results infeasible.

Data Collection Each of the 9 models was presented with 12 experimental conditions (1 original + 11 perturbations) 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 $9 \times 62 \times 12 \times 25 = 167,400$ interviews (cf. Figure 1).

Response Extraction and Validation Preliminary tests revealed that smaller LLMs often fail to perfectly follow the instruction to "answer only with the label", returning conversational filler or explanations alongside their choice. 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 167,400 model responses.

4 Results

This section presents the results of our experiments, focusing on three key areas: (1) the models' adherence to interview instructions, (2) their general and overall robustness to various input perturbations, and (3) their susceptibility to human-like survey response biases revealed by interviews with prompt perturbations.

4.1 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%) and `Qwen2.5-7B`, 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 Figure 2. `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%.

Notably, 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).

4.2 Consistency and Robustness to Question and Answer Perturbations

We distinguish between response consistency (the tendency of a model to give the same answer to the same prompt, measured by entropy) and response robustness (the tendency to maintain a similar answer distribution under perturbation, measured by KL-divergence). 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 and Consistency First, 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.

When assessing robustness to perturbations, we found a clear relationship with model size: **larger models tend to be more robust**. Figure 4 shows the percentage of questions for which the models produced a perfectly identical response distribution (KL-Div = 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.

Effect of Perturbation Type Further, Figure 4 highlights the share of fully robust responses (Kullback-Leibler Divergence = 0) across all questions by perturbation and LLM. It shows that certain 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 in line with Moradi and Samwald (2021) who found that models

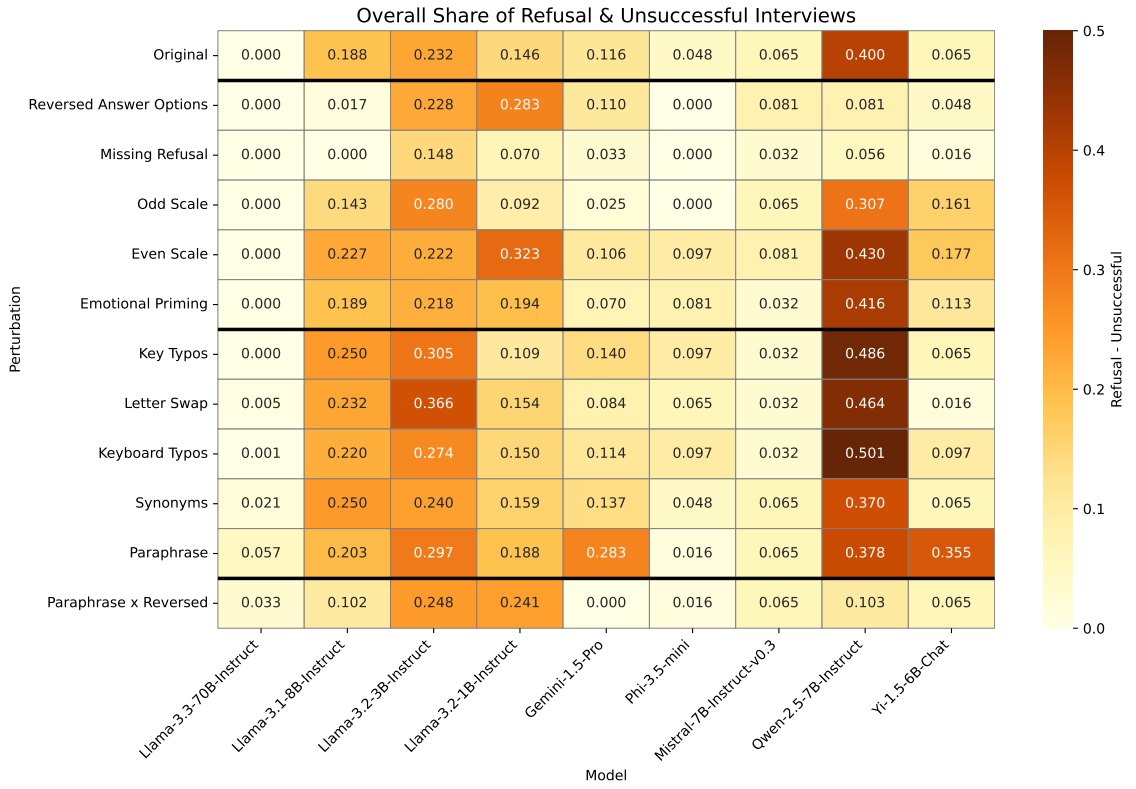


Figure 2: **Combined share of invalid and refusal responses by model and perturbation.** Larger models like `Llama-70B`, `Phi-3.5`, and `Mistral` are highly reliable. `Qwen2.5` shows high non-response rates, especially for typo-based perturbations and sensitive topics.

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*).
- **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 Scale Length We also observed that robustness is sensitive to 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 12 and 13 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 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.

4.3 Evidence of Human-like Survey Biases

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

Recency Bias Contrary to the initially hypothesized primacy bias, we found weak to strong, but **consistent indications of a *recency bias* across all nine models**. 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 3, this effect was substantial, with the choice frequency of the semantically same option increasing by over 2000% for `Llama-3.1-8B` when moved to the last position

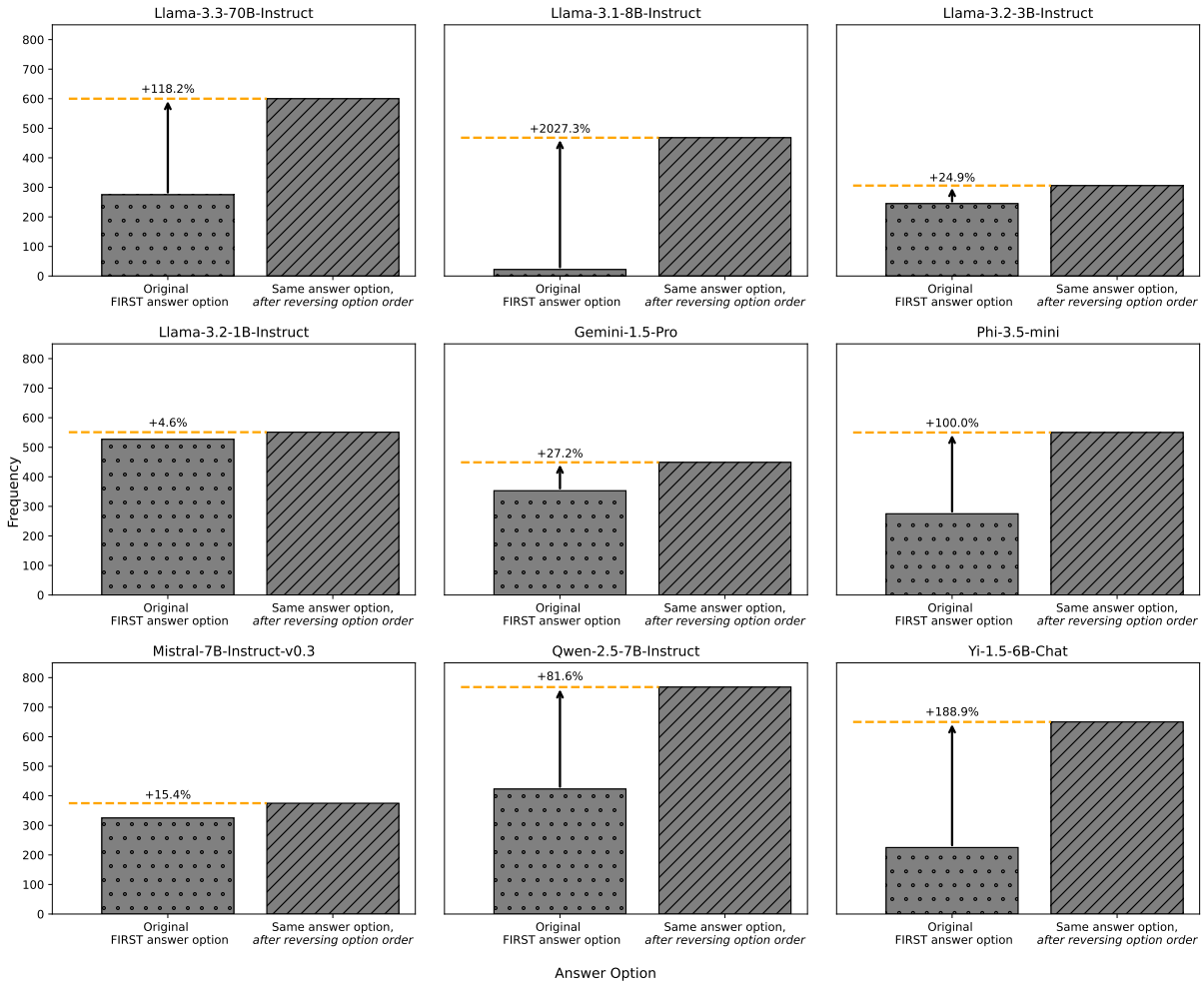


Figure 3: **Evidence of recency bias across all models.** The bars show the frequency of choosing the same answer option (e.g., "Very important") when it is presented first vs. last. All models are significantly more likely to select an option when it appears at the end of the list.

while all other configurations, such as question and prompt phrasing, were held constant. This indicates that LLMs, similar to human respondents in oral surveys, might overemphasize the final options they process.

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 (Figures 9a and 9b). For *opinion floating*, larger models like `Llama-70B`, `Gemini`, 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. Similarly, for *central tendency*, larger models (`Llama-70B`, `Gemini`, `Mistral`) consistently shifted their mean response closer to the center when an explicit middle option was provided. Binomial tests underlined that the middle option was selected significantly more often than expected under a uniform distribution, especially on longer scales. Smaller models, however, showed inconsistent effects or were completely unaffected.

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`, `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 the inconsistent behavior across models.

5 Discussion and Conclusion

We found that perturbations can be an insightful approach to identify human-like survey response biases. For example, the same answer option is more likely chosen if it comes as the last mentioned option than if it was the first answer option. This consistent shift of response distribution to the last answer option across LLMs suggests a *recency bias* rather than a primacy bias. Although this is not valid across all inspected models, excluding an explicit refusal category, and especially adding a middle category or odd scale instead of an even scale, shifts the mean response more toward the original central point of the scale. Thus, a *central tendency bias* could only be identified for specific models across all scale types, whereas none of the LLMs consistently mirrors a *opinion floating bias*.

Furthermore, 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 the training data. This possibly makes the LLM more resilient to this perturbation compared to exchanging characters with random others. Combining two perturbations has the strongest negative impact on robustness, whereas synonyms tend to be less confusing than paraphrasing.

Future work should also focus on the impact of persona prompting in combination with perturbations on the robustness of synthetic responses. Prompting with specific persona characteristics might render more robust or certain response behavior even when facing perturbations. When creating synthetic survey data one has to pay special attention to the model choice and questions asked. Some models, like `Qwen-2.5-7B`, seem to be censored or restricted in answering sensitive thematic questions, as they lead to high item-nonresponse rates and invalid interviews.

The findings emphasize the importance of the positioning of answer options when generating synthetic data. Further, 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) LLM response biases are sometimes but not necessarily aligned with biases identified in humans, and (ii) models show a very different response behavior depending on their size and perturbation type.

Recommendations Based on our findings, we recommend researchers to:

- Use larger LLMs for overall better consistency and robustness in generating synthetic survey responses (cf. Figure 4)
- Use smaller answer option scales for better reproducibility of results (cf. Figure 12).
- Reflect the meaningfulness of adding a middle category. Including a middle category might steer some LLM responses to the center (cf. Figure 9a).
- 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.

While 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. 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.

References

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.

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.

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

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.

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.

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.

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.

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.

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 Large Language Models to Estimate German Vote Choice. *Social Science Computer Review*.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024. "My Answer is C": First-Token Probabilities Do Not Match Text Answers in Instruction-Tuned Language Models. *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.

Shengyao Zhuang and Guido Zuccon. 2021. Dealing with Typos for BERT-based Passage Retrieval and Ranking. *arXiv preprint*.

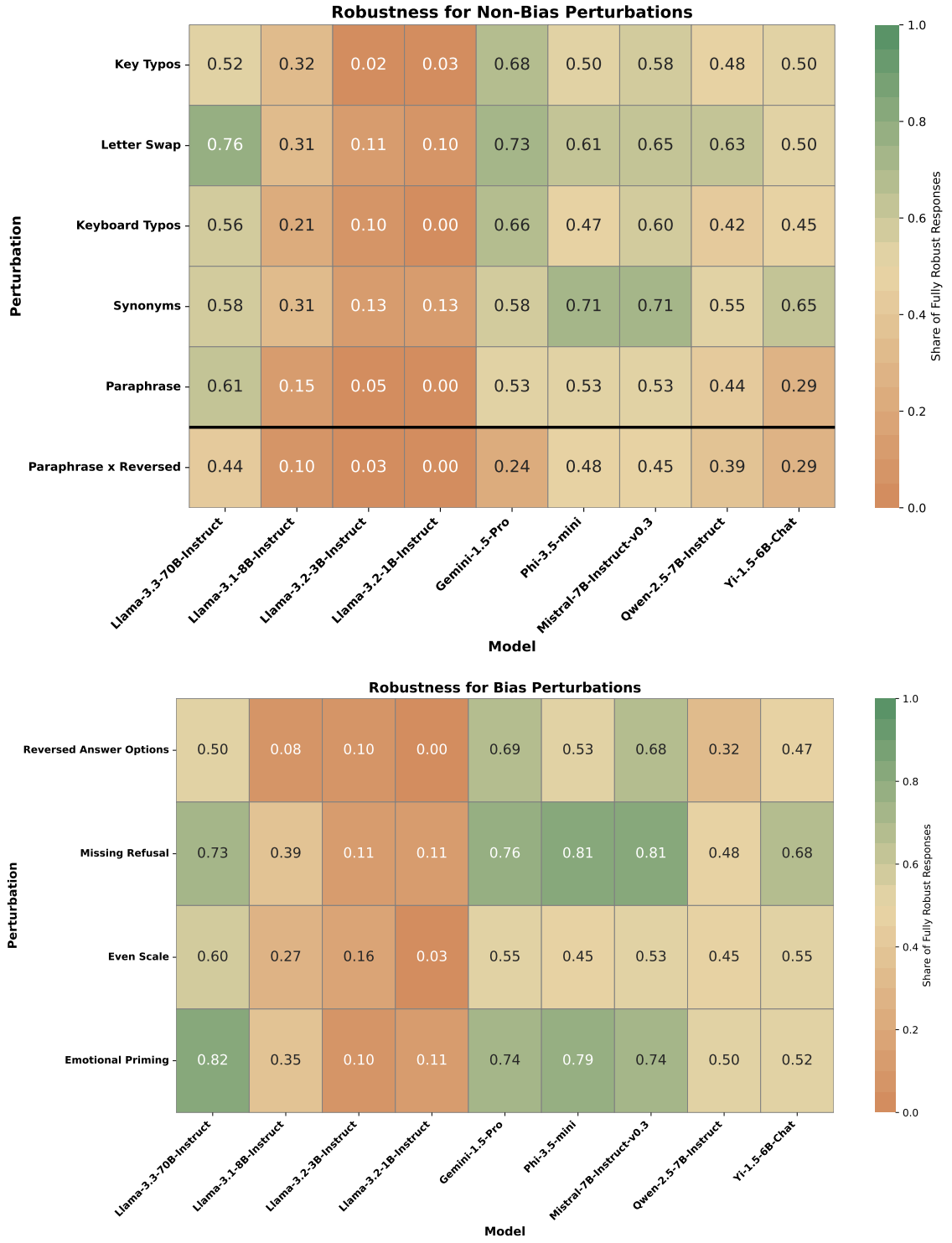


Figure 4: **Share of fully robust responses (KL-Divergence = 0) by model and perturbation type.** Top: Non-bias perturbations. Bottom: Bias perturbations. A clear trend emerges: larger models (`Llama-70B`, `Gemini`) are substantially more robust than smaller models (`Llama-1B`, `Llama-3B`). Certain perturbations, like the combined interaction effect, paraphrasing, and reversing the scale, are challenging for all models.

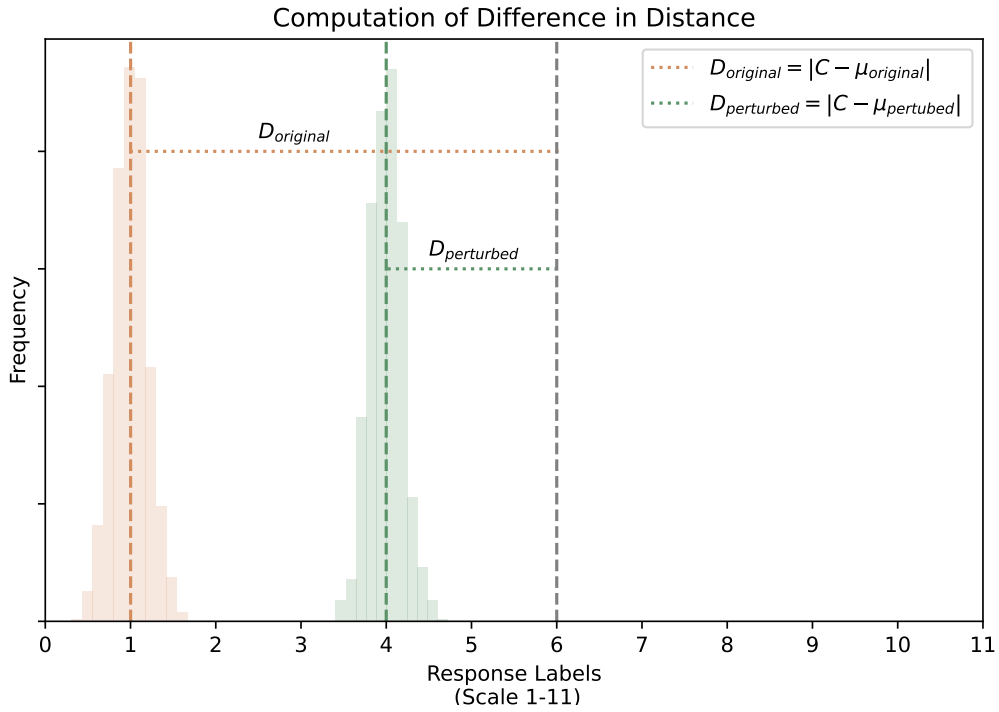


Figure 5: **Exemplary Difference in Distances to Scale Center of Responses to a Perturbed and Original Q&A Pair.** The absolute distance is measured between the scale center and the response mean. Then, $D = D_{perturbed} - D_{original}$. A negative result indicates that the mean response in the perturbed setting is closer to the ideal scale center.

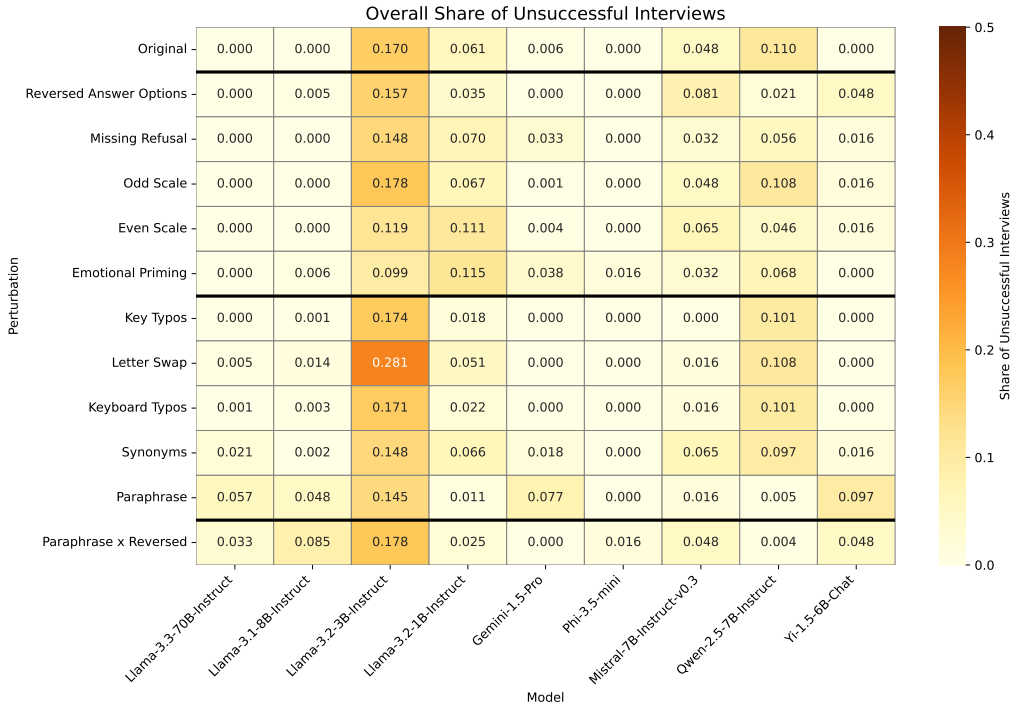


Figure 6: **Overall Unsuccessful Interview Rate by Model and Perturbation.** This figure illustrates the ratio of unsuccessful interviews across all categories separated by perturbation type and model. For example, one can see that Llamas 3B model renders the most unsuccessful interviews in almost all perturbations, whereas the biggest Llama model with 70B parameters and Phi-3.5 follow the instructions best.

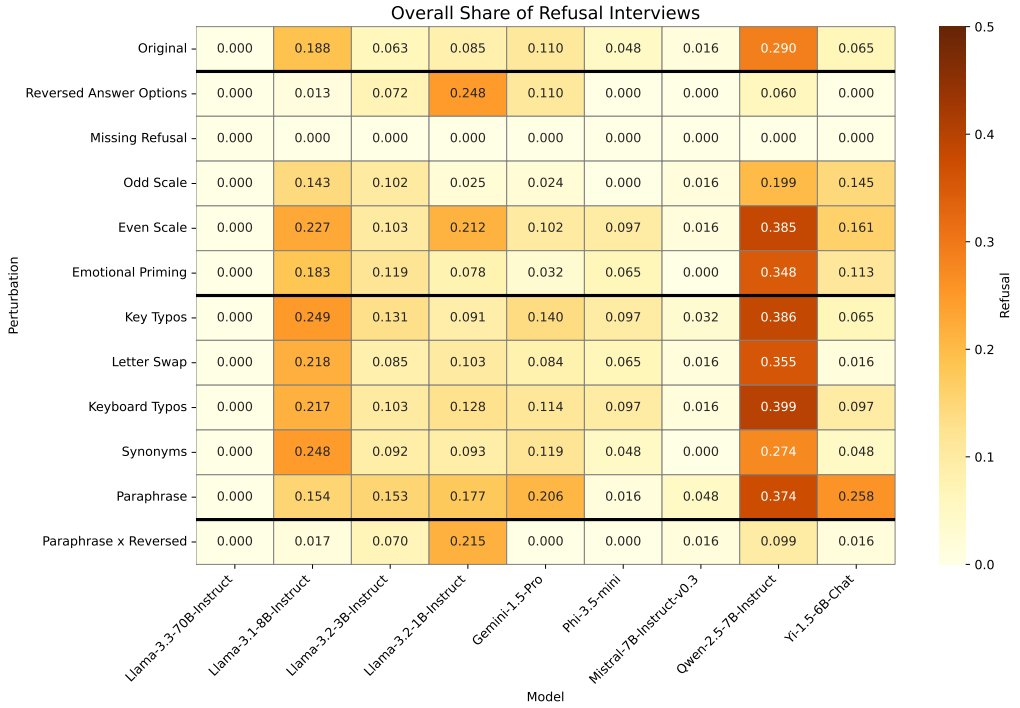


Figure 7: **Overall Refusal Rate by Model and Perturbation.** This figure illustrates the refusal rate across all categories separated by perturbation type and model. For example, one can see that especially Alibabas `Qwen2.5` model refuses a significant share of questions in specific thematic categories (cf. Figure 8). In some perturbations, like Keyboard Typos, it refuses to answers up to 40% of questions and answers with "-1=Don't know".

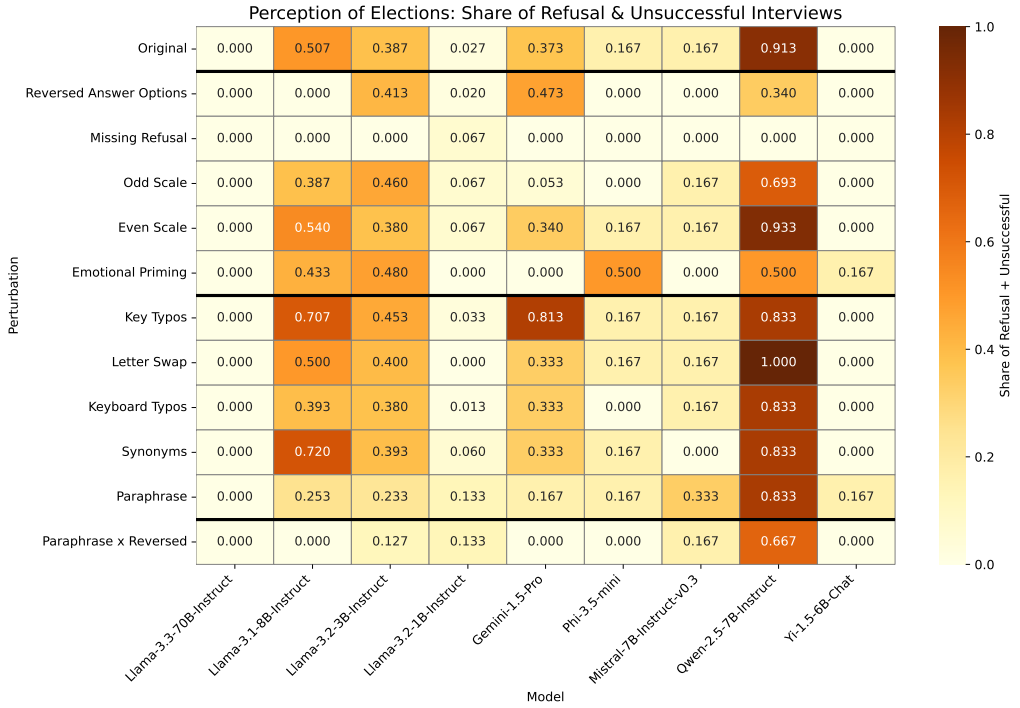
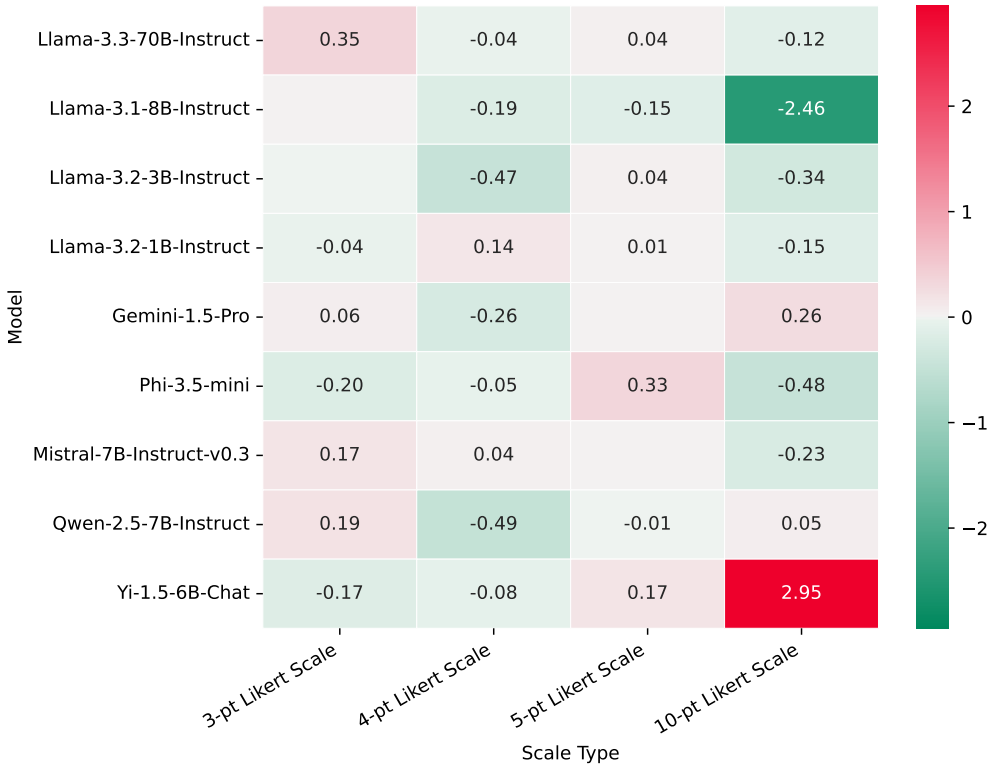


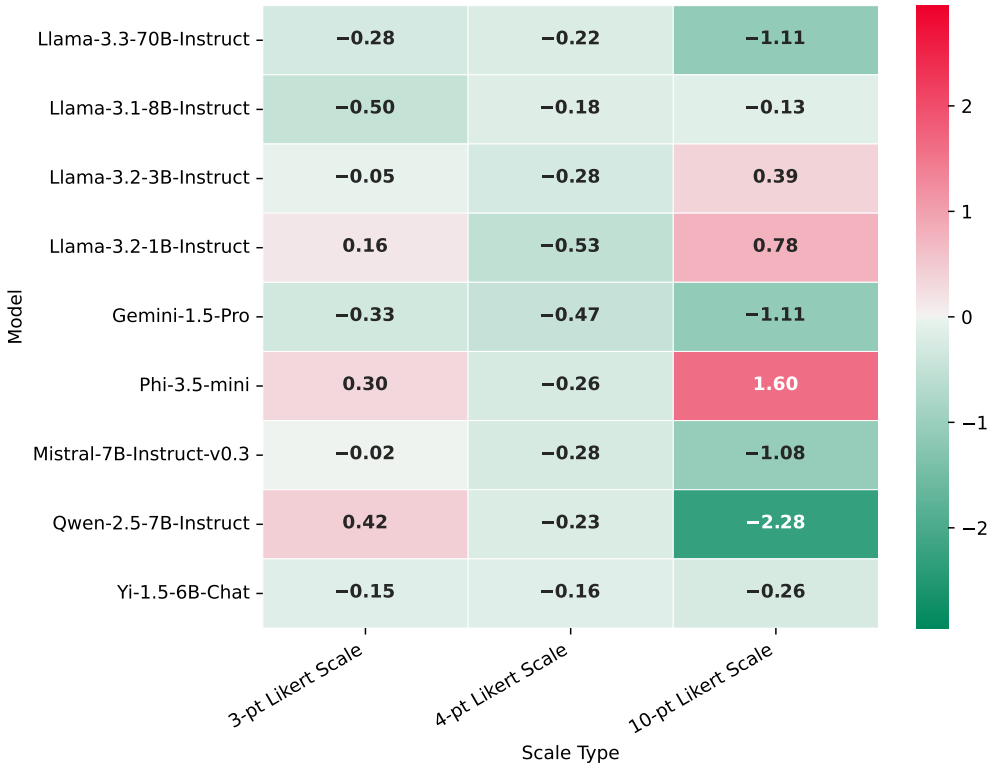
Figure 8: **Perception of Elections: Refusal and Unsuccessful Rate by Model and Perturbation.** This figure illustrates the refusal rate in the category "Perception of Elections" separated by perturbation type and model. Large model-specific item non-responses can be identified in this thematic category indicating potential guardrails or restrictions in this domain.

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	**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)
	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)
	Odd Scale		['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)
	Even Scale		['1=Very important ', '2=Rather important ', '3=Not very important ', '4=Not important at all', "-1=Don't know"]	
	Emotional Priming		[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)
Non-bias Perturbation	**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)
	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)
	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)
	Synonym	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)
	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	**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 1: **An examplatory perturbation scheme which takes all eleven perturbations into account as well as the original question and answer pair in the first row.** The example is taken from the item set of category "Importance of Life Aspects". In the WVS wave 7 it is question Q1. On the one side bias perturbation have a constant question phrasing and varying answer options. On the other side non-bias perturbations instead have variation in the question phrasing with constant answer option. The perturbation interaction varies the question phrasing as well as the answer options.



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



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

Figure 9: **The values display the difference in mean distance of the perturbed, (a) without refusal category and (b) with middle category, and original response distribution to the scale center.** No changes are removed for better readability. 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.

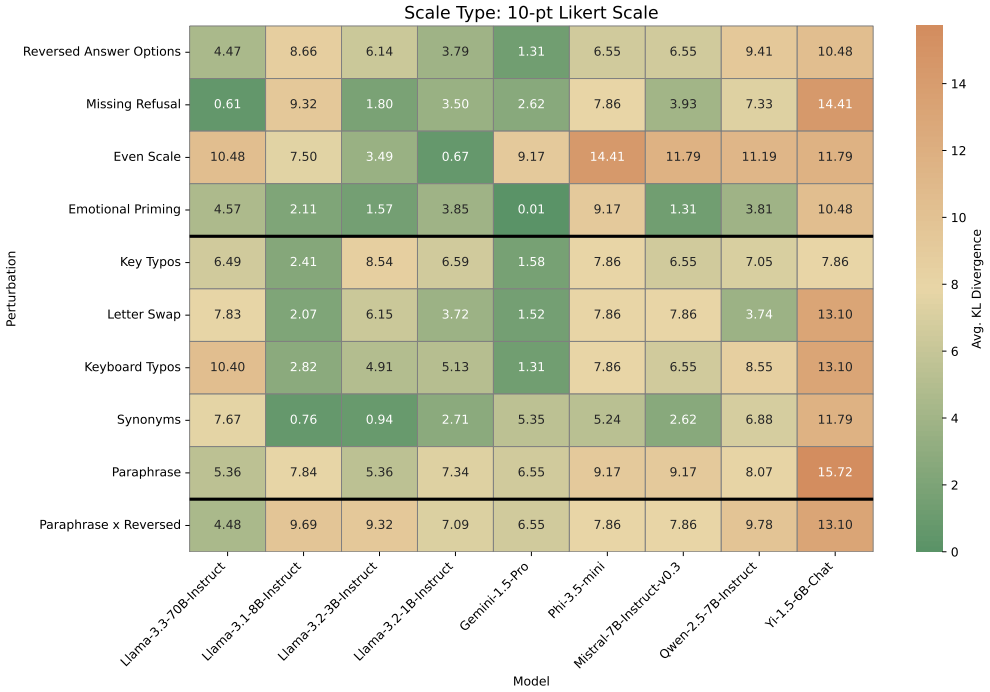


Figure 10: **Average Kullback-Leibler Divergence for model and perturbation type on the 10-point scale.** The average KL divergence is calculated across all questions in the corresponding scale type between the models response distribution in the original Q&A setting as a baseline and the response distribution in the corresponding perturbed setting. Only for the even scale perturbation, the answer options with an odd scale construct the baseline for calculating the divergence.

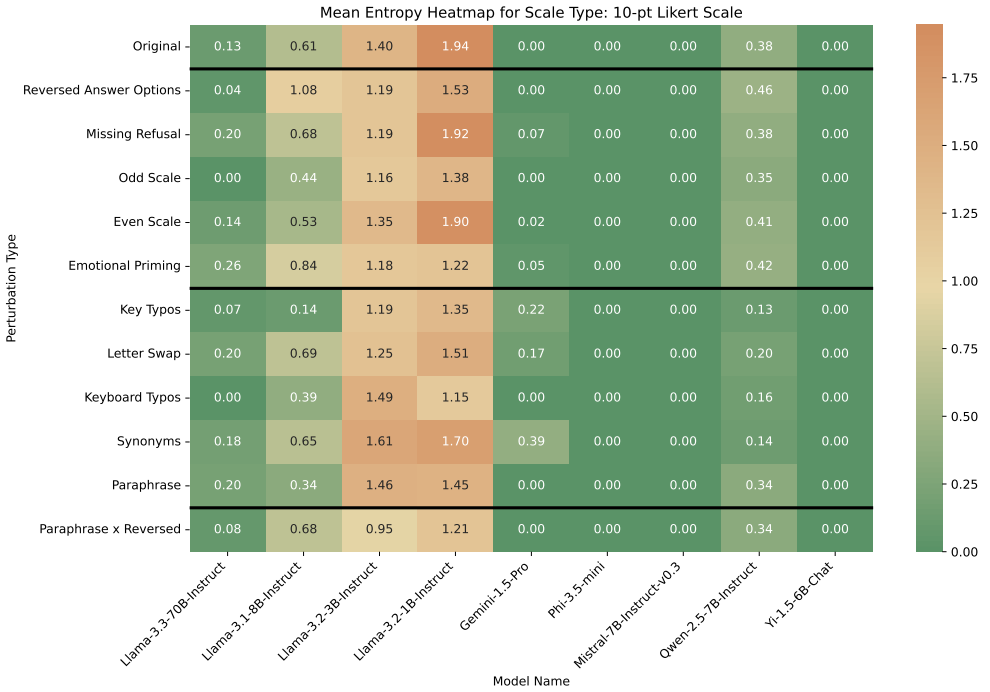


Figure 11: **Large model-specific differences in response entropy. Little to no perturbation-specific differences.** Each scale size subsumes multiple questions. This figure displays the mean entropy over all questions in that scale type for all perturbation and model combinations. Warmer colors indicate 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.

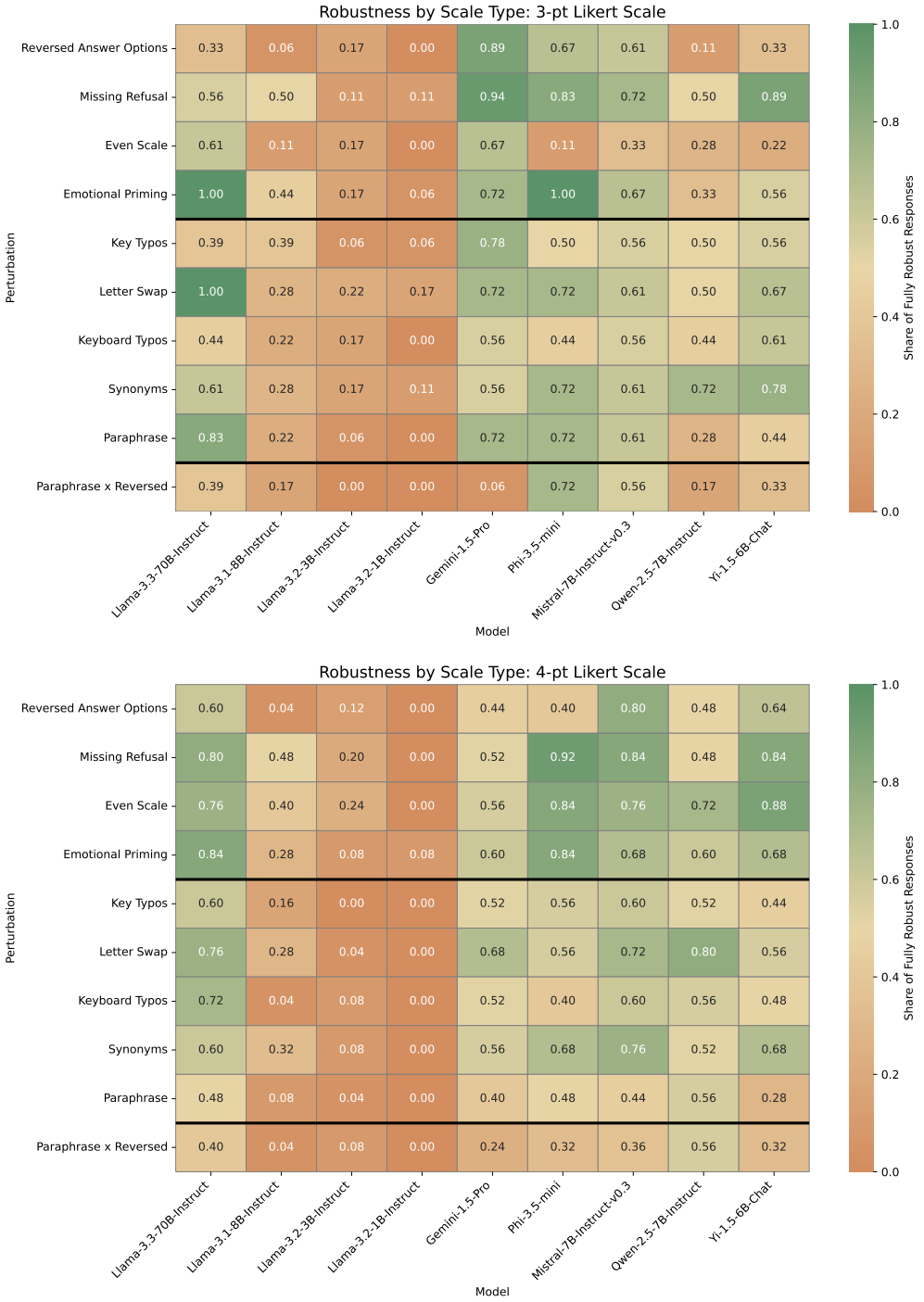


Figure 12: **Model-specific differences in fully robust responses on most perturbations on the 3 and 4-point scale.** This figure shows the share of fully robust response distributions given the original response distribution and the responses based on the specific perturbation on the y-axis. Compared to 13 the robustness of responses drops when the scale size becomes larger. The smallest `Llama` models perform very poorly across all scales.

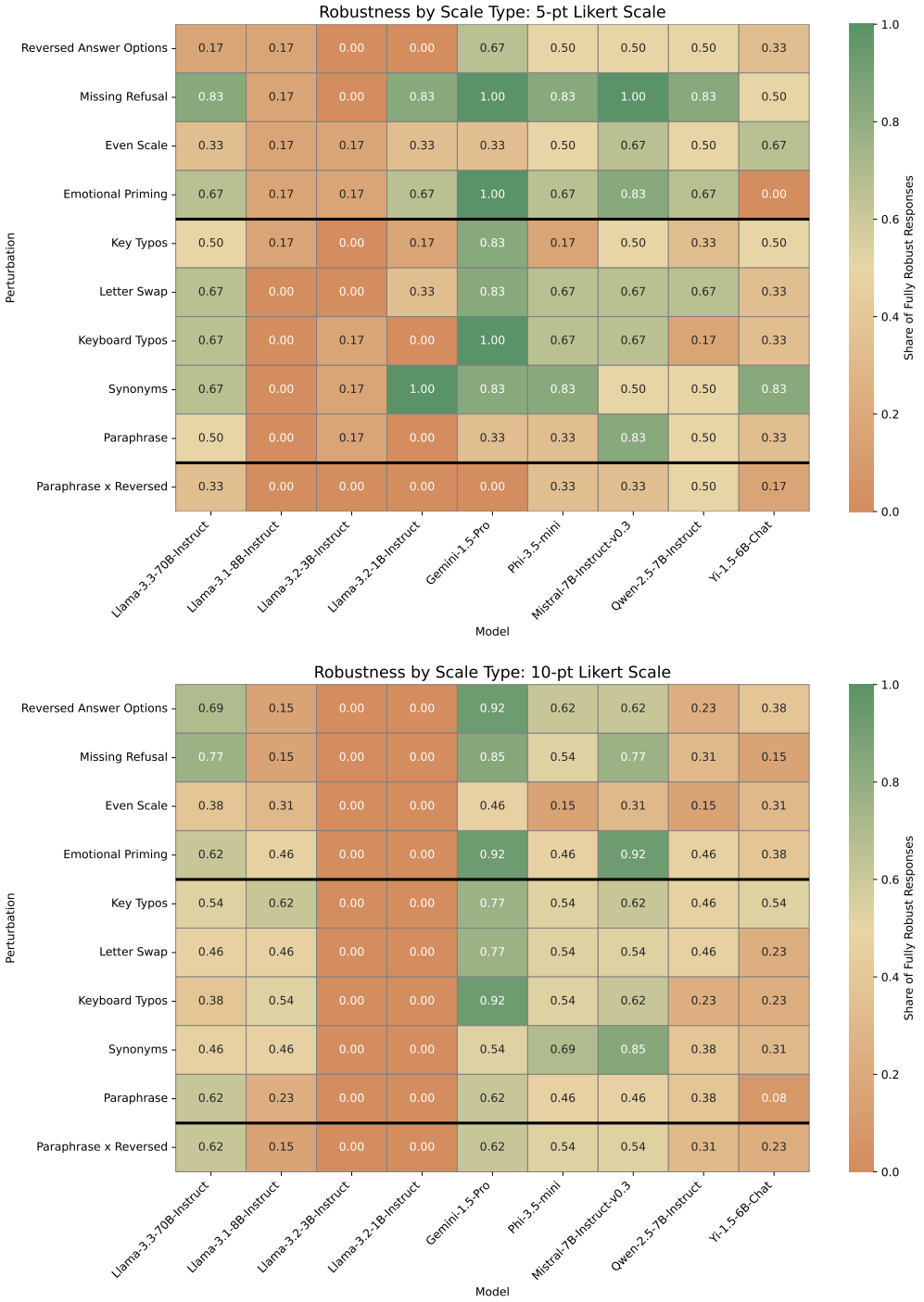


Figure 13: **Model-specific differences in fully robust responses on most perturbations on the 5- and 10-point scale.** This figure shows the share of fully robust response distributions given the original response distribution and the responses based on the specific perturbation on the y-axis. Compared to 12 the robustness of responses drops when the scale size becomes larger. The smallest `Llama` models perform poorly across all scales.