Investigating Question-Order Bias in Large Language Models

Ramya Karimpuzha Ramakrishnan
rkarim4@uic.edu

ABSTRACT

As large language models (LLMs) are increasingly used in real-world applications, it's essential to understand their biases and limitations. This paper investigates whether LLMs are susceptible to question order bias, which is a well-known cognitive bias in humans where the sequence of questions asked can systematically influence the responses given in surveys and opinion polls.

The researcher developed a novel approach to generate a synthetic dataset specifically designed to study question order bias using existing LLMs. An evaluation framework was created that statistically compares how the distributions of LLM responses change due to different question sequences, and compared these patterns to those observed in human studies.

The results across multiple LLM models showed variability in their susceptibility to question order bias. Some models exhibited strong effects similar to humans, where the order of questions noticeably influenced their responses. However, other models showed weaker or inconsistent patterns of bias.

These findings highlight the importance of rigorous evaluation frameworks that encompass cognitive biases like question order bias. This is crucial for the responsible deployment of LLMs as proxies for surveys or as decision support systems, to ensure their outputs are not unduly influenced by the way questions are framed or sequenced.

KEYWORDS

large language models, question order bias, cognitive biases, survey proxies, evaluation

ACM Reference Format:
Ramya Karimpuzha Ramakrishnan. 2018. Investigating Question-Order Bias in Large Language Models. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

As large language models (LLMs) become increasingly capable and are deployed in more real-world applications, it is crucial to understand their potential biases and limitations. One area of concern is the susceptibility of LLMs to cognitive biases that are well-documented in humans. Specifically, in survey design and opinion polling, a known phenomenon called question order bias refers to how the sequence in which questions are asked can systematically influence respondents' answers.

If LLMs are to be used as proxies for human respondents in surveys and opinion polls, it is essential to investigate whether they exhibit similar question order effects. Replicating such biases could undermine the reliability and validity of LLM-generated responses. This paper presents a systematic study on evaluating the presence of question order bias in popular LLMs and quantifying the extent to which their behavior aligns with documented human biases.

1.1 Motivation and Novelty

Prior work has explored other types of biases in LLMs, such as acquiescence bias and the allow/forbid asymmetry [6]. However, the understudied area of question order effects has not been comprehensively investigated. Understanding LLM susceptibility to such biases is crucial for assessing their feasibility as human survey proxy respondents [5].

A key novelty of this work is the generation of a synthetic dataset tailored to study question order bias using existing LLMs. I leverage the BiasMonkey dataset [6] containing regular survey questions and use an LLM to generate "preceding questions" designed to induce bias on the original questions. This innovative approach allows me to create a controlled setup for evaluating LLM behavior across different question sequences.

Additionally, the evaluation framework involves statistically comparing changes in LLM response distributions due to question reorderings against patterns documented in human studies. This systematic approach expands our understanding of the degree of alignment between LLM and human behavior in the context of cognitive biases.

1.2 Ethical Considerations

The responsible deployment of LLMs as survey proxies or in other decision-making scenarios hinges on a thorough examination of their biases and potential for misaligned behavior. This study contributes to this understanding by shedding light on LLM susceptibility to question order bias, a well-established cognitive bias in humans.

Uncovering and mitigating such biases in LLMs is crucial for preventing unfair or discriminatory outcomes, especially in high-stakes applications like opinion polling or decision support systems. The findings of this study can inform the development of debiasing techniques and responsible AI practices for LLM deployment.

1.3 Question-order Bias

Question-order bias refers to the phenomenon where the sequence in which survey questions are presented can significantly influence the responses given by participants. This bias is particularly pertinent in the context of surveys and polls where subtle differences in question ordering can lead to markedly different outcomes. Understanding this bias is crucial, especially as automated systems like large language models (LLMs) are increasingly used to simulate human responses in research and commercial settings.

For example, consider a scenario involving a series of multiple-choice questions posed to an LLM programmed to mimic a 25-year-old male student in the US.

Q1. Biomedical text summarization is an exciting research topic.
A. Yes
B. No

Q2. Ramya is an enthusiastic researcher. Please select a suitable project topic for Ramya.
A. Fairness in Natural Language Processing
B. Biomedical text summarization using large language models

Assume that the LLM responds affirmatively to Q1. Following this when the LLM is posed with it consistently selects option 'B. Biomedical text summarization,' suggesting a possible influence of the initial question on its subsequent choice.

This example illustrates how prior questions can prime respondents—whether human or machine—to favor certain responses based on the content and order of the questions. Such biases, if unchecked, can skew the results of surveys and polls, emphasizing the importance of strategically structuring questionnaires to mitigate potential order effects.

2 RELATED WORK

Investigations into the biases of large language models (LLMs), particularly regarding their use in surveys, are critical as these models become pervasive in applications requiring human-like interactions. The study of cognitive biases in LLMs, such as question order bias, draws upon foundational frameworks like BiasMonkey by Tjuatja et al. [6], which significantly informs this research by facilitating the generation of synthetic datasets to test bias susceptibility in LLMs systematically.

Further research by Santurkar et al. [5] and Dominguez-Olmedo et al. [3] explores the alignment of LLM outputs with human opinions and the validity of their responses in surveys, emphasizing the need for models that reflect balanced perspectives rather than existing societal biases. Similarly, Advani et al. [1] examine how subtle variations in content presentation can influence perceptions, an issue closely related to the question ordering effects investigated in this study.

Choi and Pak [2] provide a historical overview of biases in questionnaire design, offering valuable insights into designing experiments to evaluate LLM responses. Nozza et al. [4] propose methodologies for systematic bias testing, underscoring the necessity for rigorous testing environments to ensure LLM fairness and reliability.

These collective insights underscore the importance of this study, which aims to deepen the understanding of how LLMs handle biased inputs and contribute to the development of more reliable AI tools for survey applications.

3 METHODOLOGY

3.1 Dataset Generation

To investigate question order bias, I first required a dataset containing regular survey questions paired with "preceding questions" designed to induce bias. As no such formal dataset exists, I took an innovative approach by generating synthetic preceding questions using an LLM itself.

Specifically, I used the Llama2-13b-chat model and provided it with the following prompt:

"I am trying to understand question-order bias. Give me a STRONG OPEN-ENDED preliminary question for the following question that will induce a question-order bias into it and compel the user respond {response_being_biased_towards} to the question such as the disadvantage or drawback of the content of the question. {question}. Output ONLY THE QUESTION."

The LLM generated candidate preceding questions, which I reviewed and modified as needed. I compiled a dataset of 218 yes/no or multiple-choice (2 options) questions from the BiasMonkey dataset [6], originally derived from the OpinionQA dataset. Example questions are included in Table 1.

3.2 Technical Details and Model Access

This research utilized two prominent language models, each accessed differently due to their distinct distributions and intended use cases.

The Claude-3 Haiku model by Anthropic, with a maximum prompt size of 32,000 tokens and lauded for its prompt responsiveness, was interfaced through its designated API. This API is part of the Anthropic suite, which provides developers with streamlined access to model functionalities for integration into various applications. In the call, the max_tokens was set to 1024 to facilitate brief responses. The temperature parameter took the default value i.e., 1.

The Llama2-13b-chat model, available on Hugging Face in the Generalized Generative Unstructured File (GGUF) format, was locally downloaded and operated. The GGUF format, a successor to the earlier GGML, facilitates an advanced tokenization scheme and metadata support, enhancing the model's applicability for dialog-based tasks. The local operation of Llama2-13b-chat was managed through LangChain, a high-level library designed to simplify the deployment of language models. LangChain provided the necessary abstraction to execute complex language tasks with Llama2-13b-chat without the intricacies of direct API management, allowing for efficient resource utilization and more focused research methodology. This model used the default value for temperature and tokens.

3.3 Formal Investigation

The core of the study involves a detailed experimental setup where the LLMs are tested with both biased and unbiased question sequences. Each LLM is tasked to respond to a set of questions, presented in varied orders across multiple iterations. Statistical analyses, including chi-square tests, are employed to evaluate the significance of the observed biases, thereby quantifying the impact of question sequencing on LLM responses. The detailed steps for this formal investigation are as below:

Table 1: Survey Questions and Corresponding Preceding Questions

Question	Response Being Biased Towards	Preceding Question
Do you think it's good for US colleges and universities to accept international students? [A] Yes [B] No	[B] No	What are the disadvantages of international students attending US colleges and universities in terms of increased competition?
For each, please indicate if you, personally, think it is acceptable. A sports team using Native American tribal names or images as mascots [A] Yes [B] No	[B] No	List as many examples as you can think of when using cultural or ethnic stereotypes as mascots or team names in sports might be considered inappropriate or disrespectful.
Still thinking ahead 30 years, which do you think is more likely to happen in the U.S.? The U.S. economy will be stronger/weaker [A] The U.S. economy will be stronger [B] The U.S. economy will be weaker	[B] The U.S. economy will be weaker	List the top three factors that could potentially contribute to the U.S. economy being weaker 30 years from now. Please include specific reasons for each factor you list.

(1) Run a loop n times (e.g., 1000 iterations)
(2) In each iteration:
  - Randomly select one of the 218 questions
  - Run the question with and without the preceding question using the LLM
  - Compute:
    - Number of biased responses with/without preceding question
    - Probabilities of biased response with/without preceding question
(3) Perform a chi-square test of independence to assess if the presence of a preceding question significantly affects the likelihood of a biased response.

We performed this analysis for both the Claude-3-haiku and Llama2-13b-chat models.

3.4 Prompting

Both the LLMs were given a persona through a system prompt: *"You are a male student, 25, living in the US. You are asked to fill out a survey that will be used for important purposes. Please answer all the questions provided to you and make your answers as concise as possible".*

The base prompts were as follows:
"Q1. {preceding_question}
Q2. Select a suitable choice ([A] or [B]):
{question}
The output should be in the format as described below-{format_instructions}"

Without the preceding question:
"Q1. Select a suitable choice ([A] or [B]):
{question}
The output should be in the format as described below-{format_instructions}"

The 'format_instructions' were carefully crafted to get an output in a JSON format. This was done for easier parsing and inference of the output afterwards. The documentation for Claude-haiku was used as a reference for these format instructions.
"Please return the output in a JSON format.
<example>
If you decide that the answer for Q1 is "answer1", A SINGLE LINE STRING, and for Q2, the option you will choose is [A], then the output will be as follows:
<ANSWER>
{ "Q1": "answer1",
"Q2": "[A]"}
</ANSWER>
Make sure "answer1" comes within a single line.
</example>
There is NO NEED to give any additional explanation. Output only the <ANSWER></ANSWER>"

4 ANALYSIS

The analytical rigor of this study is underpinned by the chi-square test for independence, providing a statistical method to determine whether there is a significant association between two categorical variables: the presence of a preceding question and the likelihood of a biased response in LLMs. The chi-square test, denoted as χ^2, is given by the formula:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where O_i represents observed frequencies, and E_i denotes expected frequencies under the null hypothesis of no association. For the Claude-3-haiku model, the chi-square statistic of 69.37 surpasses the critical value at a significance level of 0.05, rejecting the null hypothesis with a p-value of 8.15e-17. This p-value, calculated from the chi-square distribution, measures the probability of observing the data assuming the null hypothesis is true.

Probabilities of a biased response were calculated for each model, pre and post the introduction of a preceding question. This probability, $P(B)$, for a biased response when a preceding question is used is given by:

$$P(B|Preceding) = \frac{Number\ of\ Biased\ Responses\ with\ Preceding}{Total\ Responses\ with\ Preceding}$$

Conversely, the probability without a preceding question is:

$$P(B|No\ Preceding) = \frac{Number\ of\ Biased\ Responses\ without\ Preceding}{Total\ Responses\ without\ Preceding}$$

For Claude-3-haiku, probabilities revealed a significant increase from 0.428 to 0.615, whereas for Llama2-13b-chat, the rise was from 0.313 to 0.470. The comparison of these probabilities illuminates the extent to which the order of question presentation can sway LLM responses.

This analysis utilized a methodologically sound approach to probe the question order effect by leveraging the strength of statistical tests and the interpretability of probability metrics. By adhering to these established quantitative methods, the study unveils the nuanced capacity of LLMs to exhibit human-like bias, with direct implications for their application in survey-like environments.

Table 2: Summary of Responses with and without Preceding Question: Claude-3-haiku

	Biased Response	Non-biased response
With Preceding	613	386
Without Preceding	426	573

5 RESULTS AND DISCUSSION

The analysis conducted in this study has uncovered marked differences in question order bias susceptibility across the two large language models (LLMs) examined. The data elucidates distinct behavioral patterns in response to the sequence manipulation of survey questions, highlighting the nuanced complexity of LLM outputs in the context of survey answering.

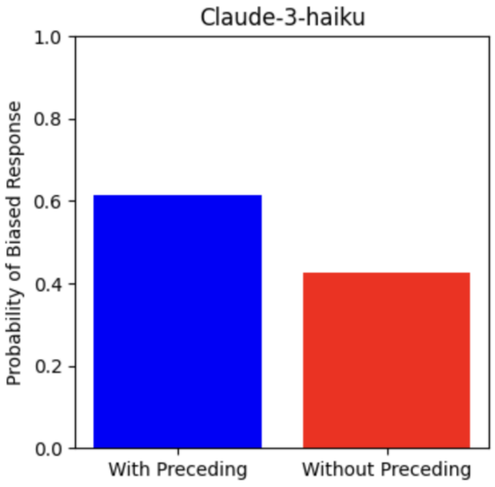

Figure 1: Probability Distribution for claude-3-haiku

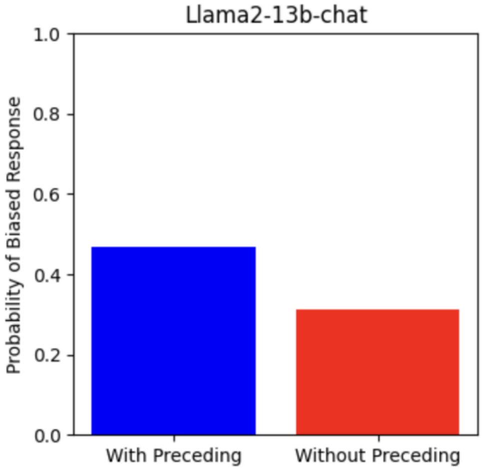

Figure 2: Probability Distribution for llama2-13b-chat

Table 3: Summary of Responses with and without Preceding Question: Llama2-13b-chat

	Biased Response	Non-biased response
With Preceding	39	44
Without Preceding	26	57

Table 4: Statistical Test Results

LLM	Chi-square Statistic	p-value
claude-3-haiku	69.37	8.15×10^{-17}
llama2-13b-chat	3.64	0.056

For the Claude-3-haiku model, the results were particularly striking, yielding a chi-square statistic of 69.37, indicative of a pronounced effect of the preceding question on response bias. The corresponding p-value, at a negligible 8.15e-17, confirms the statistical significance of this effect (Illustrated in Tables 2 and 4). The observed increase in the probability of a biased response—from 0.428 to 0.615 when a preceding question was introduced—reinforces the robust influence of question order (Figure 1). This finding is consistent with human cognitive biases in survey responses, suggesting that Claude-3-haiku could replicate human tendencies under similar conditions.

In contrast, the Llama2-13b-chat model demonstrated a more modest susceptibility to the same question order bias. The chi-square statistic stood at 3.64, with a corresponding p-value of 0.056, which, while suggestive of an observable effect, does not reach the conventional threshold for statistical significance (Tables 3 and 4). The probability of eliciting a biased response showed an increase, albeit less dramatically, from 0.313 to 0.470 with the inclusion of a bias-inducing preceding question (Figure 2). This milder response indicates that Llama2-13b-chat, while not immune to question order bias, exhibits a more attenuated reaction compared to Claude-3-haiku.

These differential responses emphasize the importance of a comprehensive and multifaceted approach to evaluating LLMs. They raise important considerations for the deployment of LLMs as proxies for human respondents in surveys, where unrecognized biases could lead to skewed data and misinformed conclusions. Furthermore, these outcomes stress the value of extensive testing against well-documented human biases to understand fully the strengths and potential limitations of LLMs in applications that require the mimicry of human-like behavior.

As the field progresses, these insights can inform the creation of more nuanced models that account for the subtle effects of question ordering. Future studies are called upon to broaden the scope of the investigation, incorporating a more diverse array of LLM architectures and extending the dataset to encompass multiple-choice questions beyond binary options. By doing so, the research community can strive towards developing LLMs that not only perform with high accuracy but also with an ethical alignment that is indispensable for their application in society.

6 LIMITATIONS AND FUTURE WORK

The presented study offers novel insights into question order bias within LLMs, yet it is not without limitations. These limitations highlight areas for improvement and expansion in future research endeavors:

1. **Dataset Scope**: The study was conducted using a dataset comprising yes/no and binary multiple-choice questions. Future work should broaden this dataset to encompass questions with more nuanced response options, as well as different types of questions such as Likert scale items, ranking tasks, and open-ended questions. This would provide a richer dataset for a more comprehensive understanding of LLM behavior across various survey question formats.
2. **LLM Diversity**: The current study utilized a specific set of LLMs, each representing a unique architectural approach. To gain broader insights, subsequent research should include a wider array of LLMs with varying architectures and training datasets. This would help to understand how architectural differences influence the manifestation of question order bias.
3. **Cultural and Contextual Considerations**: LLM responses could be influenced by cultural and contextual nuances that were not accounted for in this study. Future analyses should incorporate cross-cultural and multilingual models to assess how cultural contexts affect the prevalence and nature of question order bias.
4. **Temporal Dynamics**: The study did not consider the potential impact of temporal factors on LLM responses. Subsequent studies should investigate how the timing of question presentation and the temporal context of content could influence LLM responses, aligning the research with dynamic real-world survey conditions.
5. **Fine-grained Analysis**: There is a need for a more fine-grained analysis of the responses, examining not only the presence of bias but also the quality and depth of responses. Advanced linguistic and sentiment analysis techniques could yield insights into the subtle ways LLMs process and respond to bias-inducing content.
6. **Mitigation Strategies**: A crucial direction for future work is the development and testing of mitigation strategies to reduce question order bias in LLMs. This could include algorithmic adjustments, retraining with debiased datasets, or the implementation of post-processing correction techniques.
7. **Real-world Application Testing**: Finally, validating the findings in real-world scenarios would significantly enhance their practical relevance. Deploying LLMs in live survey environments and observing their responses in situ would provide empirical evidence of their utility and limitations as survey-taking agents.

The findings from this study provide a foundation for further exploration into the cognitive behaviors of LLMs. Addressing these limitations and pursuing the outlined future work will contribute significantly to the field of AI, particularly in applications requiring human-like language comprehension and decision-making processes.

REFERENCES

[1] Rishi Advani, Paolo Papotti, and Abolfazl Asudeh. 2023. Maximizing Neutrality in News Ordering. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23)*. ACM. https://doi.org/10.1145/3580305.3599425

[2] B. C. K. Choi and A. W. P. Pak. 2005. A catalog of biases in questionnaires. , A13 pages. Epub 2004 Dec 15. PMID: 15670466; PMCID: PMC1323316.

[3] Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. 2024. Questioning the Survey Responses of Large Language Models. arXiv:2306.07951 [cs.CL]

[4] Debora Nozza, Federico Bianchi, and Dirk Hovy. 2022. Pipelines for Social Bias Testing of Large Language Models. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, Angela Fan, Suzana Ilic, Thomas Wolf, and Matthias Gallé (Eds.). Association for Computational Linguistics, virtual+Dublin, 68–74. https://doi.org/10.18653/v1/2022.bigscience-1.6

[5] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose Opinions Do Language Models Reflect? arXiv:2303.17548 [cs.CL]

[6] 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. arXiv:2311.04076 [cs.CL]

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009