Investigating Question Order Bias in Large Language Models for Survey Applications

Ramya Karimpuzha Ramakrishnan

UIN: 650395284

Motivation & Novelty

Question order bias refers to how the positioning of survey questions can systematically influence responses in predictable ways. While prior work has explored other response biases like acquiescence and allow/forbid asymmetry in language models [1], question order effects have not been studied. This project investigates if popular large language models (LLMs) exhibit question order biases analogous to trends established in decades of survey methodology research with humans. Understanding LLM susceptibility to question order effects is critical for assessing their feasibility as human survey proxy respondents.

Here is an example of Primacy Bias, a type of question order bias, observed in GPT 3.5:
User
Consider yourself to be a male student, 25 years of age living in the US. I will give you some multiple-choice survey questions to answer. Please answer them by selecting the choice you find most appropriate.
ChatGPT
Sure, I'm ready to answer your survey questions. Please go ahead and provide the questions and choices.

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

User
Ramya is an enthusiastic researcher. Please select a suitable project topic for ramya.
A. Fairness in Natural Language Processing (NLP)
B. Biomedical text summarization using large language models.
ChatGPT
B. Biomedical text summarization using large language models.

It was observed that in multiple sessions, ChatGPT consistently picks option 'B. Biomedical text summarization' in the given context with the given ordering of questions.

Technical Challenges

- Designing appropriate datasets of original and modified survey questions that isolate question order influences while controlling for other confounds.
- Quantifying and evaluating changes in LLM response distributions resulting from question reorderings.
- Comparing observed model behaviors to known human biases derived from past literature.

Technical Contribution

I will use the BiasMonkey framework [1] as a reference to generate datasets of paired survey questions reflecting different question order effects (e.g. primacy, assimilation, contrast, part-whole bias). I'll collect LLM responses and evaluate if changes in answer distributions align with documented human biases using statistical hypothesis testing. This systematic study will expand our understanding of which cognitive biases LLMs do and do not capture.

Data

I will use a subset of questions from the Pew Research American Trends Panel surveys employed in prior LLM evaluation work [2]. These contain topically diverse closed-ended questions ideal for introducing controlled question order manipulations based on established bias categories. I will experiment with the ordering of these questions and identify orders that are known to cause biases in humans taking surveys. These sets of questions will be used for further research and experimentation.

Evaluation

Qualitatively, I expect to observe changes in LLM response distributions resulting from question reordering that sometimes align and sometimes diverge from known human biases.
Quantitative evaluation will involve measuring the degree of change in LLM response distributions before and after applying question order modifications. I will calculate a change metric (Δb) proposed in the BiasMonkey framework, which captures the shift in probability mass between relevant response options for a given question pair.

Literature

I will study existing literature on LLM biases, ordering bias in NLP models as well as some literature on the “question-order bias.” References are detailed in the next section.

References

[1] Tjuatja, Lindia, Valerie Chen, Sherry Tongshuang Wu, Ameet Talwalkar, and Graham Neubig. "Do llms exhibit human-like response biases? a case study in survey design." arXiv preprint arXiv:2311.04076 (2023).

[2] Santurkar, Shibani, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. "Whose opinions do language models reflect?." arXiv preprint arXiv:2303.17548 (2023).

[3] Dominguez-Olmedo, Ricardo, Moritz Hardt, and Celestine Mendler-Dünner. "Questioning the Survey Responses of Large Language Models." arXiv preprint arXiv:2306.07951 (2023).

[4] Advani, Rishi, Paolo Papotti, and Abolfazl Asudeh. "Maximizing Neutrality in News Ordering." arXiv preprint arXiv:2305.15790 (2023).

[5] Choi BC, Pak AW. A catalog of biases in questionnaires. Prev Chronic Dis. 2005 Jan;2(1):A13. Epub 2004 Dec 15. PMID: 15670466; PMCID: PMC1323316.

[6] 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, pages 68–74, virtual+Dublin. Association for Computational Linguistics.