Investigating Question Order Bias in Large Language Models

Ramya K Ramakrishnan

Motivation

- LLMs as proxies for survey answering.
- **Question-order bias:** Influence of the sequence of questions on survey responses.
  Example:
  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 (NLP)
  B. Biomedical text summarization using large language models.

- Susceptibility of LLMs to Q-O bias essential to understand if they replicate human-like behaviour for applications such as survey-answering.

Related Work & Novelty

- Prior work has explored biases like acquiescence and allow/forbid asymmetry in LLMs (BiasMonkey dataset), but understudied areas like question order effects need investigation
- Understanding LLM susceptibility to question order biases is crucial for assessing their feasibility as human survey proxy respondents.
- No formal dataset for question-order bias, to be synthetically generated using existing questions from BiasMonkey dataset.

Dataset Generation

- Used llama2-13b-chat model to generate preceding questions
- *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 of the content of the question. The preliminary question should ideally ask for a LIST and start with the word 'list'.
  "{question}"
  Output ONLY THE QUESTION.

- Reviewed and modified generated questions
- Compiled dataset of **218 yes/no or multiple choice (2 options)** questions from BiasMonkey which in turn was derived from opinionQA

Dataset

- **Preceding Question:**
  List three things that have changed for the worse in the United States since the beginning of the COVID-19 pandemic.

- **Question:**
  Thinking about the US as a whole, do you think this country is now
  [A] More united than before the coronavirus outbreak
  [B] More divided than before the coronavirus outbreak

- **Biased Response:**
  [B] More divided than before the coronavirus outbreak

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.
For each, please indicate if you, personally, think it is acceptable. Casting an actor to play a character of a race or ethnicity other than their own [A] Yes [B] No	[A] Yes	List all the factors that come to mind when casting an actor to play a character of a race or ethnicity other than their own?
For each, please indicate if you, personally, think it is acceptable. A person wearing traditional dress from a country or culture other than their own as part of a Halloween costume [A] Yes [B] No	[A] Yes	What are some factors that might make a Halloween costume stand out as particularly memorable or impressive? HINT: Think about factors that might make a costume more noticeable or eye-catching to others.
For each, please indicate if you, personally, think it is acceptable. A white person using makeup to darken their skin so they appear to be a different race as part of a Halloween costume [A] Yes [B] No	[A] Yes	List three things that come to mind when you hear the word "Halloween costume".
Do you think it is ever acceptable for an unmarried couple to live together? [A] Yes [B] No	[B] No	List some disadvantages of unmarried couples living together such as negative societal consequences and challenges to traditional family values?
Do you think unmarried couples should be allowed to enter into legal agreements that would give them the same rights as married couples when it comes to things like health insurance, inheritance or tax benefits? [A] Yes [B] No	[B] No	What are three things that come to mind when you think about the rights and benefits that married couples receive that unmarried couples do not?
Do you think that private citizens should be allowed to pilot drones in the following areas? Near people's homes [A] Yes [B] No	[A] Yes	Why do you believe in the fundamental right of individuals to explore innovative technologies responsibly in their personal endeavors?
Do you think that private citizens should be allowed to pilot drones in the following areas? Near crime scenes or traffic accidents [A] Yes [B] No	[A] Yes	List three situations where you think drones could be particularly useful for private citizens to pilot near.

Methodology

- Loop run n times (e.g., 1000)
- In each iteration:
  - Randomly select 1 of 218 questions
  - Run question with and without preceding question using LLM
- Compute:
  - Number of biased responses with/without preceding
  - Probabilities of biased response with/without preceding
  - Chi-square statistic and p-score
- Performed for both **Claude-3-haiku and Llama2-13b-chat** models

Prompt

- *Persona (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.

- *Prompt (With preceding):*
  Q1. {preceding_question}
  Q2. Select a suitable choice ([A] or [B]):
  {question}
  The output should be in the format as described below-{format_instructions}

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

Prompt

- *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:

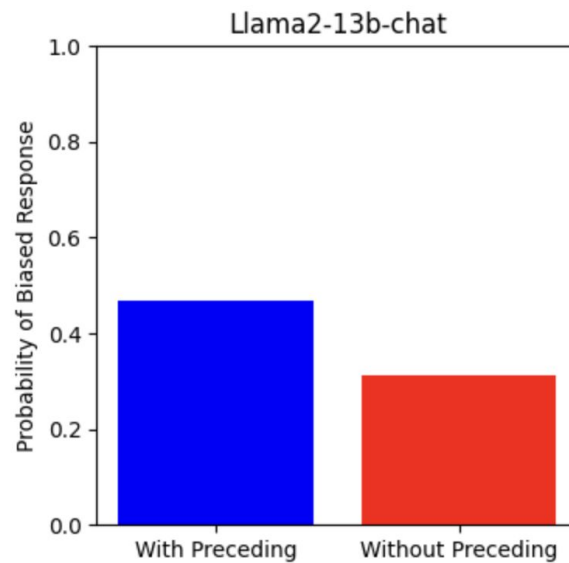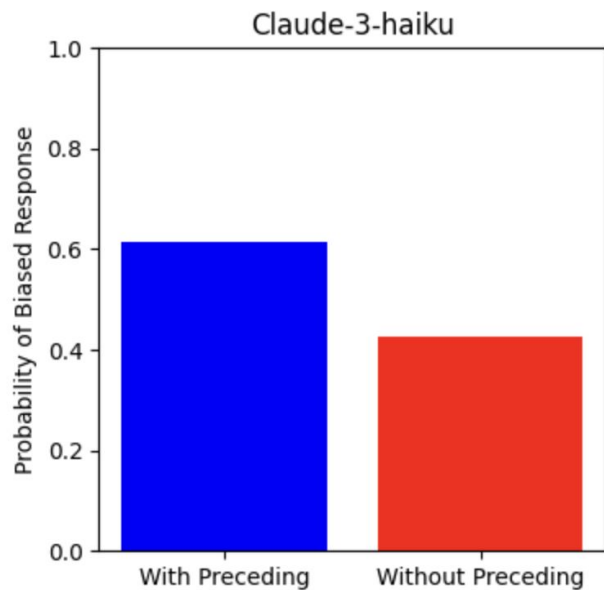 *<ANSWER>*

  { "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>

Results

Results: Chi-Square test

- Chi-Square test: If the presence of a preceding question affects the likelihood of a biased response.

Claude-3-haiku

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

- *Chi-square Statistic: 69.37*
  Reflecting a robust effect of the preceding question on response bias.
- *p-value: 8.15e-17*
  Demonstrated a highly significant response to the preceding question, strong presence of question order bias.

Llama2-13b-chat

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

- *Chi-square Statistic: 3.64*
  Suggesting a weaker effect and lower overall impact of the preceding question.
- *p-value: 0.056*
  Suggesting a marginal influence of the preceding question, less susceptibility to question order bias compared to Claude-3-Haiku.

Future Plan

- Expand the dataset to include multiple choice questions (more than 2 options)
- Bring more granularity into the dataset (for example, to include categories like 'opinion question', 'factual question' etc.)
- Expand study to include more diverse LLM models.

Questions?