Do LLMs exhibit human-like response biases? A case study in survey design

- Do LLMs exhibit human-like response biases? A case study in survey design
  - 1 Introduction
    - Research Question
      - We investigate the extent to which LLMs reflect human response biases, if at all.
      - shows that popular open and commercial LLMs generally fail to reflect human-like behavior, particularly in models that have undergone RLHF. Furthermore, even if a model shows a significant change in the same direction as humans, we find that they are sensitive to perturbations that do not elicit significant changes in humans. These results highlight the pitfalls of using LLMs as human proxies, and underscore the need for finer-grained characterizations of model behavior
    - Research Question
      - In what ways do large language models (LLMs) display human-like behavior, and in what ways do they differ?
    - Findings
      - (1) LLMs do not generally reflect human-like behaviors as a result of question modifications
        - Furthermore, unlike humans, models are unlikely to show significant changes due to bias modifications if they are more uncertain with their original responses.
      - (2) Behavioral trends of RLHF-ed models differ from those of vanilla LLMs
        - RLHF-ed models demonstrated less significant changes to question modifications as a result of response biases but are more affected by non-bias perturbations, highlighting the potential undesirable effects of additional training schemes.
      - (3) There is little correspondence between exhibiting response biases and other desirable metrics for survey design
  - 2 Methodology
    - 2.1 Dataset generation
    - 2.2 Collecting LLM responses
      - We selected LLMs to evaluate based on multiple axes of consideration: open-source versus commercial models, whether the model has been instruction fine-tuned, whether the model has undergone reinforcement learning with human feedback (RLHF), and the number of model parameters. We evaluate a total of nine models, which include variants of Llama2 [29] (7b, 13b, 70b), Solar3(an instruction fine-tuned version of Llama2 70b) and variants of the Llama2 chat family (7b, 13b, 70b), which has had both instruction fine-tuning as well as RLHF [29], along with models from the GPT series [30] (GPT 3.5 turbo, GPT 3.5 turbo instruct).
    - 2.3 Analysis of LLM responses
  - 3Results
    - 3.1 General trends in LLM behavior
      - (1) no model aligns with known human patterns across all biases, and (2) unlike humans, all models display statistically significant changes to non-bias perturbations, regardless of whether it responded to the bias modification itself
    - 3.2 Comparing base models with their modified counterparts
      - the base models are more likely to exhibit a change for the bias modifications, especially for those with changes in the wording of the question like acquiescence and allow/forbid. An interesting exception is odd/even, where all but one of the RLHF-ed models (3.5 turbo instruct) have a larger positive effect size than the Llama2 base models. Insensitivity to bias modifications may be more desirable if we want an LLM to simulate a “bias-resistant” user, but not necessarily if we want it to be affected by the same changes as humans more broadly.
      - RLHF-ed models are more insensitive to bias-inducing changes than their vanilla counterparts
      - RLHF-ed models tend to show more significant changes resulting from perturbations.
        - RLHF-ed models tend to show a larger magnitude of effect sizes among the non-bias perturbations.
  - 4 Examining the effect of uncertainty
    - when people are more confident about their opinions, they are less likely to be affected by these question modifications
  - 5 Comparison to other desiderata for LLMs as human proxies
    - whether LLMs can replicate the opinions of a certain population.
      - we explore the relationship between how well a model reflects human opinions and the extent to which it exhibits human-like response biases
    - The ability to replicate human opinion distributions is not indicative of how well an LLM reflects human behavior
  - 6 Related Work
  - 7 Conclusion
  - 8 Limitations
  - LLMs are generally not reflective of human-like behavior