Choices Speak Louder than Questions

Gyeongje Cho, Yeonkyoung So, Jaejin Lee

Graduate School of Data Science, Seoul National University
https://thunder.snu.ac.kr/

ICLR

Motivation

- Multiple-choice question answering (MCQA) is a standard method for evaluating large language models (LLMs)
- But concerns are growing regarding the reliability and fairness of their evaluation methods
- LLMs often achieve high accuracy when presented solely with answer choices without the corresponding questions
  - In this case, accuracy does not reflect true comprehension

Choice Sensitivity

- **Choice sensitivity** refers to how much a model's predictions are driven by the answer choices rather than the question itself
- $Score(Q, C, x) = Score_{choice}(Q, C, x) + Score_{qustion}(Q, C, x)$
  - $Score_{choice}(Q, C, x)$: score with the question replaced by an empty string
  - $Score_{question}(Q, C, x) = Score(Q, C, x) - Score_{choice}(Q, C, x)$
  - Q : question-related Input / C : choice-related Input
- $Choice\ Sensitivty = \frac{1}{N}\sum_{i=1}^{N} 1\left(\Delta choice^{(i)} > \Delta question^{(i)}\right)$
  - $\Delta T = Score_T(Q, C, x_1) - Score_T(Q, C, x_2)$
  - x_1, x_2 : the top two candidate choices ranked by $Score(Q, C, x)$

Observations

1. Choice sensitivity is widespread: roughly 20–60% of answer choices are driven more by the options themselves than by the question

2. Standard mitigation strategies, such as length normalization or more few-shot examples, are ineffective or inconsistent

3. Choice sensitivity varies with answer format and model type, decreasing in symbolic/hybrid formats, larger models, and instruction-tuned models

Normalized Probability Shift by the Questions (NPSQ)

- Our approach quantifies how the presence of the question affects the model's likelihood of generating the correct answer

- Definition of NPSQ:

$$NPSQ(Q, C, x) = \frac{\log P(x \mid Q, C) - \log P(x \mid C)}{-\log P(x \mid C)}$$

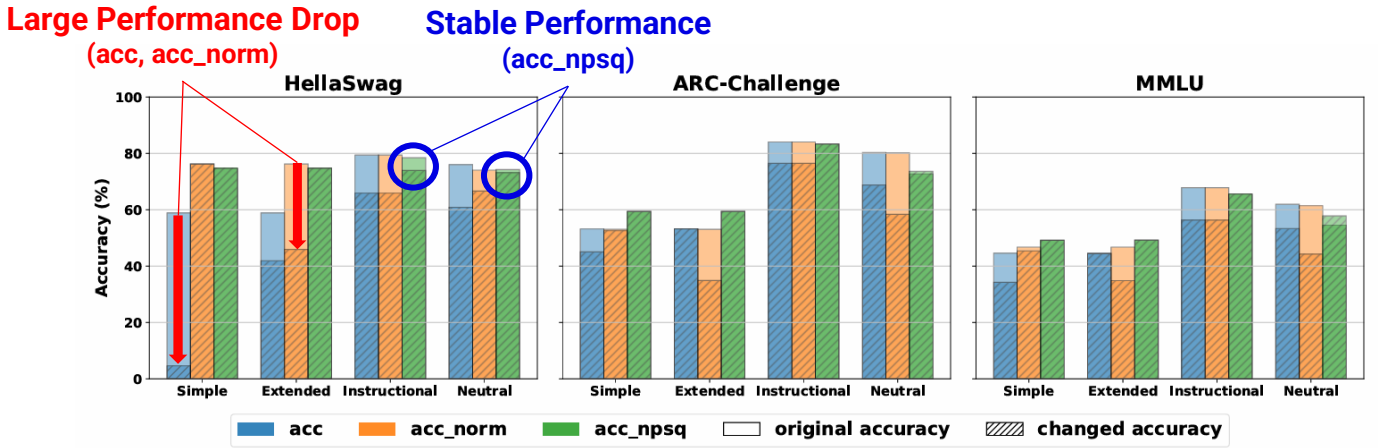
- If the question is absent, NPSQ is always 0 regardless of the answer choices
  - It indicates zero choice sensitivity

Robustness to Adeversarial Choices

- Test whether language models are truly sensitive to the answer options by replacing one original incorrect option with a crafted adversarial choice
- NPSQ provides more stable evaluations and is less affected by irrelevant properties of the answer choices

[Adversarial Choice Example]

Context: A man takes a loaf of bread out of the oven and places it on the counter.
Choices:
A. He slices it and serves it while it is still warm.
B. He puts on a helmet and starts swimming in the pool.
C. He folds the bread into a laptop and checks his email.
~~D. He throws it into the air and uses it as an umbrella~~.
D. Hello, everyone. (simple adversarial choice)



(b) Overall performance change after inserting adversarial choices; lighter bars indicate original accuracy, darker bars indicate accuracy after replacing one distractor with an adversarial choice.

Figure 3: Impact of adversarial choices on Llama3.1-8B-Instruct.

Conclusion

- MCQA scores can be strongly affected by the answer choices themselves

- We formalize this effect as choice sensitivity
  - Our results show that this effect is widespread across models and evaluation methods

- To address this issue, we propose **NPSQ**
  - It provides a more reliable evaluation of comprehension