Plausibly Problematic Questions in Multiple-Choice Benchmarks for Commonsense Reasoning

Shramay Palta, Nishant Balepur, Peter Rankel,
Sarah Wiegreffe, Marine Carpuat and Rachel Rudinger

Problem Setting

- Commonsense situations can admit multiple plausible answers.
- MCQ benchmarks need one gold answer.
- Is the gold answer always the most plausible answer?

Methodology

- Collect plausibility judgments on a 5-point Likert Scale for each (q, c_i) tuple for a question q with choices c_1, ...c_n.
- Collect best answer choice judgements.

What we present

- For 250 questions from Social IQa and CommonsenseQA:
  - 5000 Likert scale based human (crowdsourced)plausibility judgements.
  - 1530 best answer judgements.

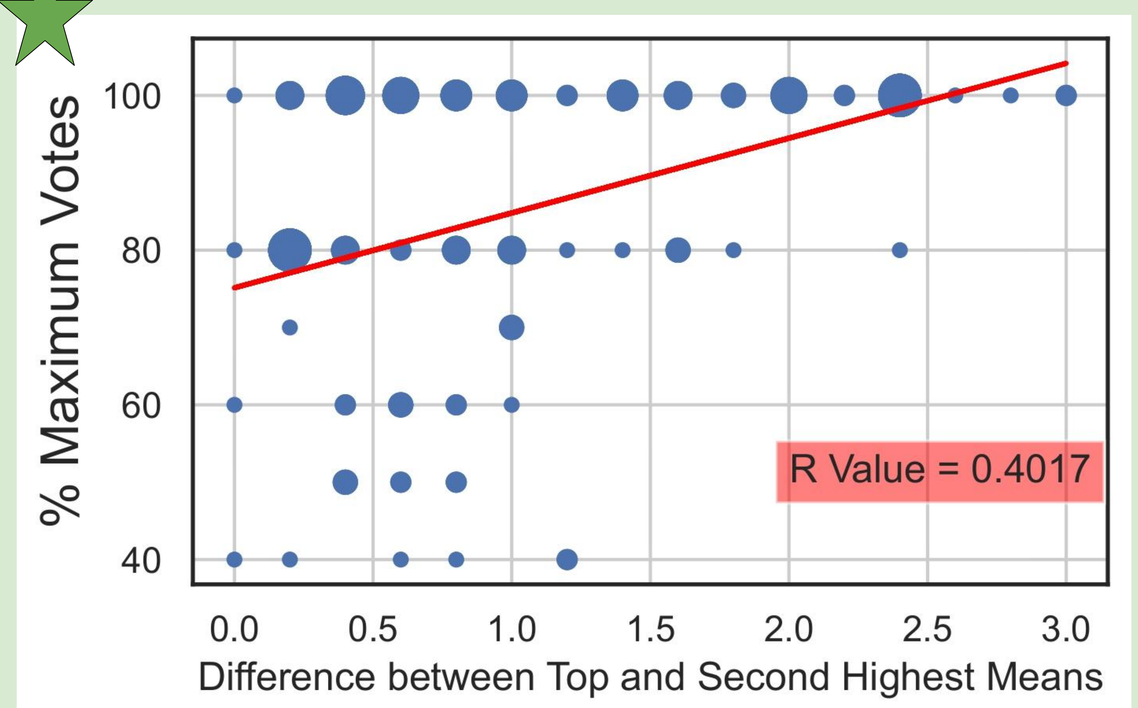
Key Takeaways

- Gold answer ≠ most plausible answer in over **20%** of the cases. → "plausibly problematic" questions. (Example on top right.)
- Qualitative analysis of these questions reveals a high prevalence of issues like question ambiguity and semantic mismatch between question and answer choices.
- MCQs with a small difference in plausibility ratings of most- and second-most plausible answer choice reflect low agreement on the best answer choice setting.
- Answer-level plausibility is a reliable way to identify problematic commonsense MCQ test items.
- LLMs have low accuracy on these 'plausibly problematic' instances.

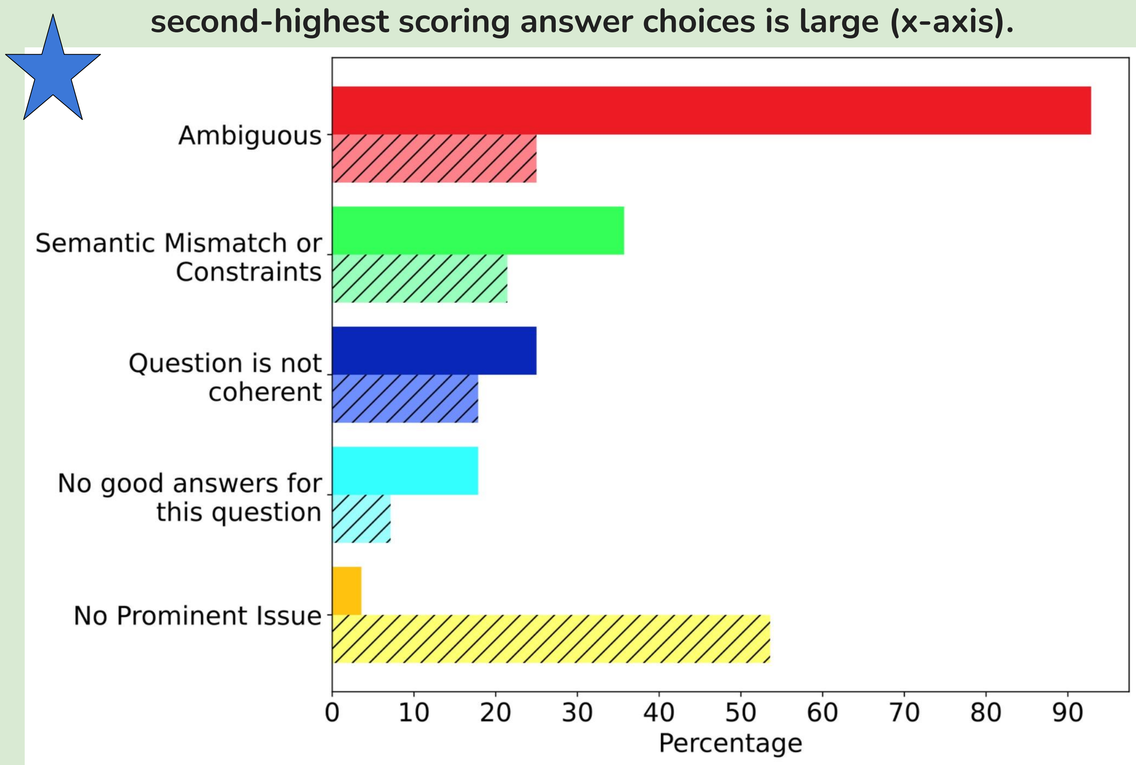
Context: Ash redeemed themselves after retaking the test they failed.
Question: How will Ash feel as a result?

AnswerA: relieved	👱 : 5, 2, 5, 5, 4	*(4.2)*
AnswerB: **accomplished**	👱 : 4, 2, 5, 2, 5	*(3.6)*
AnswerC: proud	👱 : 4, 5, 5, 5, 5	*(4.8)*

An example of a "plausibly problematic" MCQ item from SocialIQa shown with our collected plausibility ratings. The dataset gold answer (accomplished) did not receive the highest average plausibility rating from our annotators.



Annotators are more likely to agree on one correct answer (y-axis) when the gap in plausibility scores between highest- and second-highest scoring answer choices is large (x-axis).



Frequency of different issue types on the 'plausibly problematic' (solid) and non-problematic questions (hatched) from Social IQa.

References

[1] Social IQa: Commonsense Reasoning about Social Interactions](https://aclanthology.org/D19-1454) (Sap et al., EMNLP-IJCNLP 2019
[2] CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge](https://aclanthology.org/N19-1421) (Talmor et al., NAACL 2019