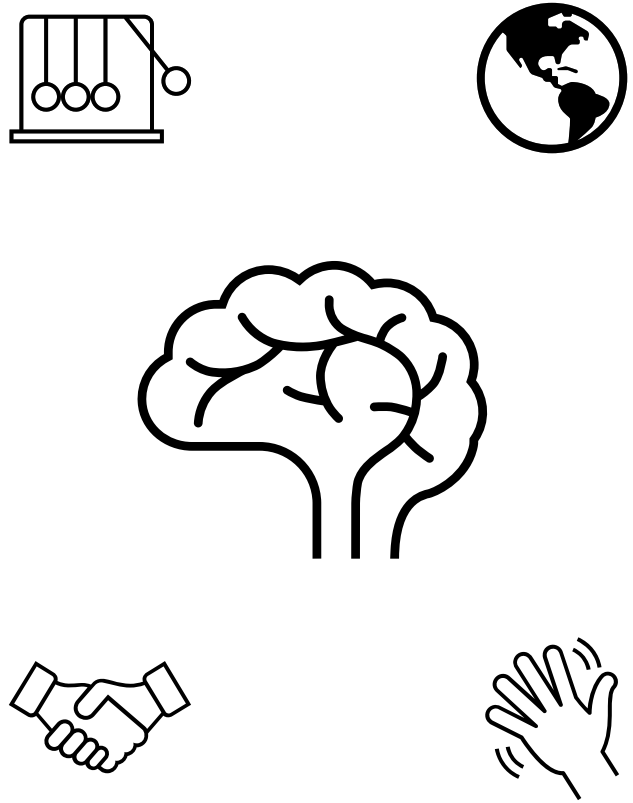
GRADED JUDGMENTS OF PLAUSIBILITY IN COMMONSENSE REASONING

Shramay Palta
University of Maryland

Motivation

- Humans have been acquiring knowledge as far back as we can trace history!
- Incorporating AI with this knowledge is important!
- Broadly two types of knowledge
  - *Factual knowledge*
  - *Commonsense knowledge*



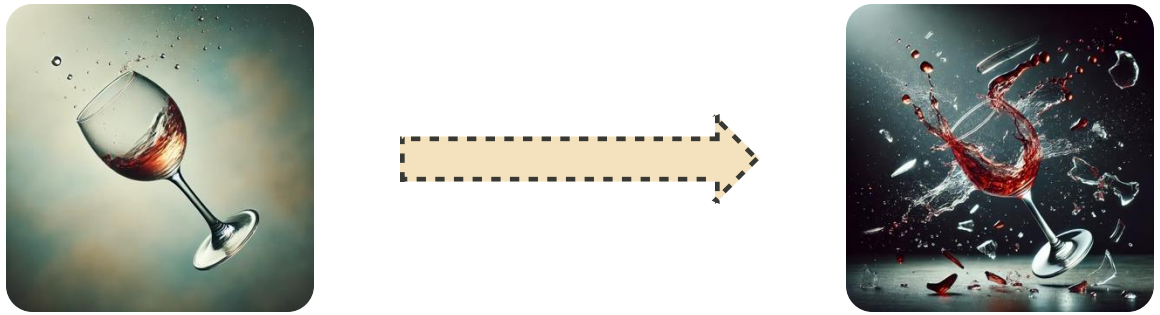
Motivation

- What is the key difference between these two types of knowledge?
  - *Factual knowledge is mostly acquired through formal instruction or training*

- But what about commonsense knowledge?
  - *Implicit and involves developing intuitions*

What is commonsense reasoning?

- Commonsense reasoning is the ability to reason about everyday scenarios.



What is commonsense reasoning?

- Commonsense reasoning is the ability to reason about everyday scenarios.

Evaluating Knowledge Acquisition

- Important to test whether humans or AI have learnt something correctly.
- How do we do that?

A familiar setup?

How many sides does a septagon have?

A) 8
B) 6
C) 7
D) 5

In what year did India gain her independence from the British?

A) 1947
B) 1946
C) 1950
D) 1949

MCQ Evaluation

- A question is presented with a fixed set of answer choices
  - *One correct answer choice (gold label)*
  - *Rest are incorrect (distractors)*

MCQ Evaluation

- One of the most widely used formats for evaluating knowledge acquisition.
  - *Humans: SAT, GRE, GMAT*

MCQ Evaluation

- One of the most widely used formats for evaluating knowledge acquisition.
  - *Humans: SAT, GRE, GMAT*
  - *AI: MMLU, CommonsenseQA, ARC*

MEASURING MASSIVE MULTITASK LANGUAGE UNDERSTANDING

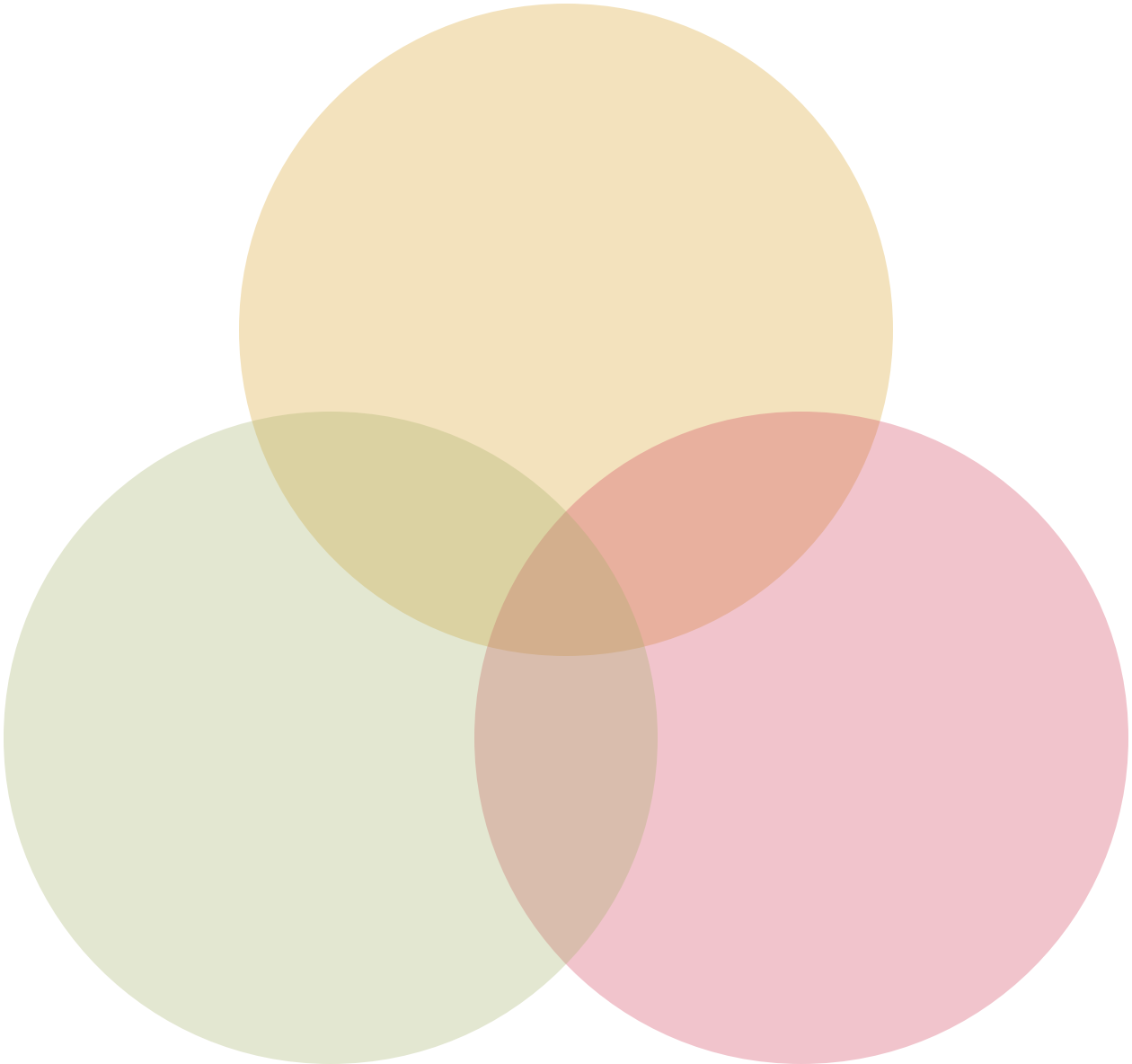
Dan Hendrycks UC Berkeley
Collin Burns Columbia University
Steven Basart UChicago
Andy Zou UC Berkeley
Mantas Mazeika UIUC
Dawn Song UC Berkeley
Jacob Steinhardt UC Berkeley

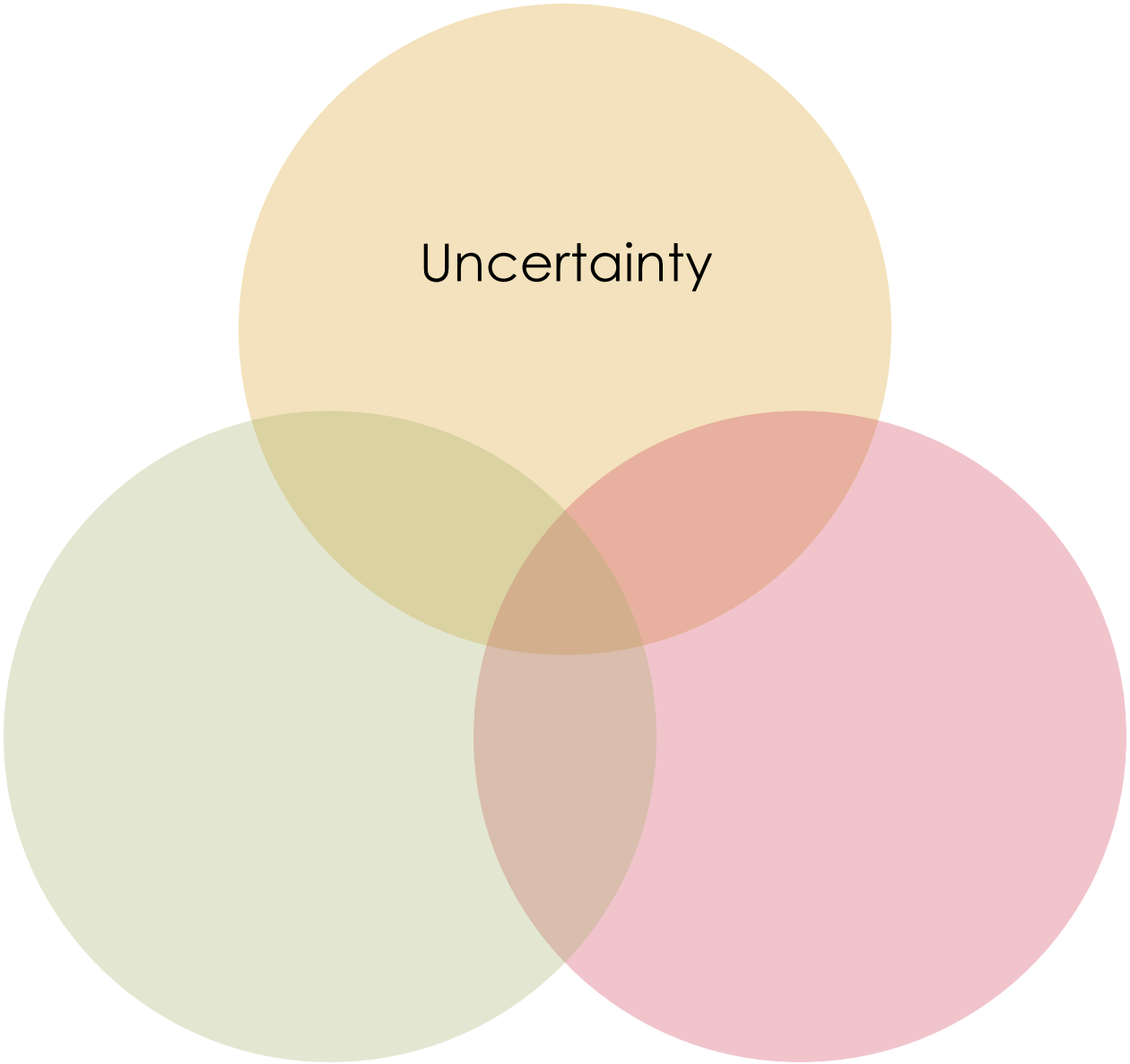
Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge

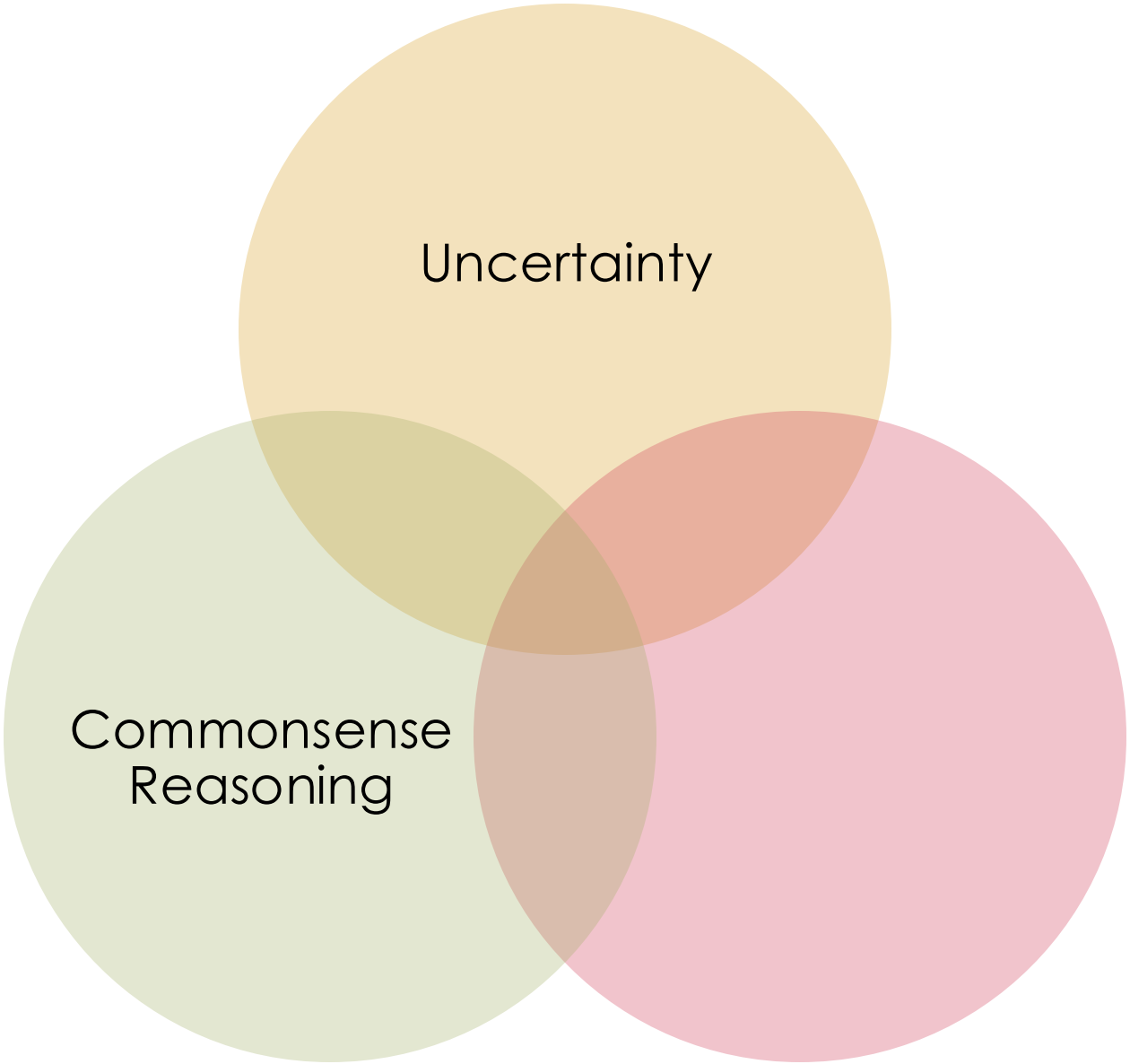
Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, Oyvind Tafjord

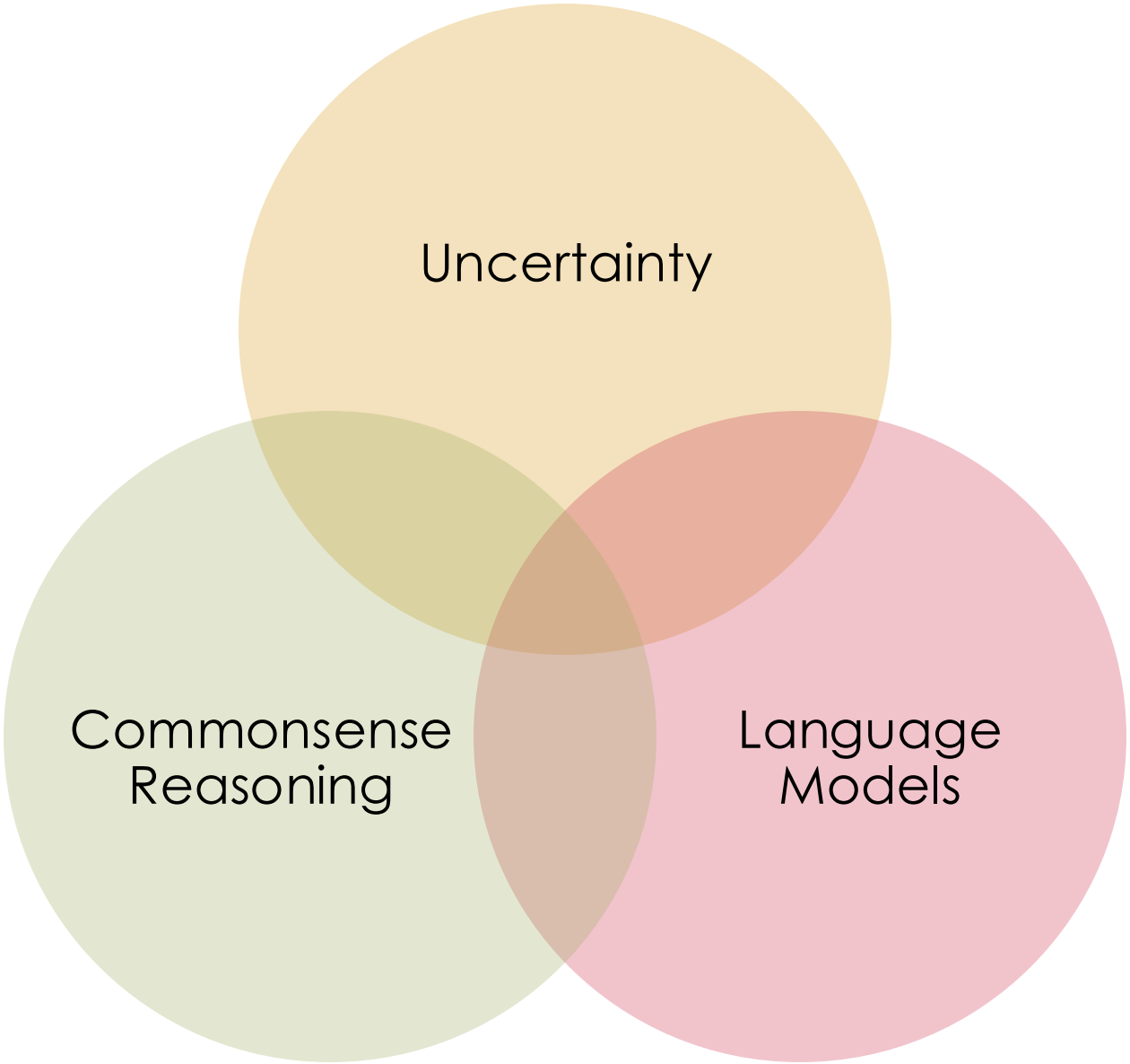
COMMONSENSEQA: A Question Answering Challenge Targeting Commonsense Knowledge

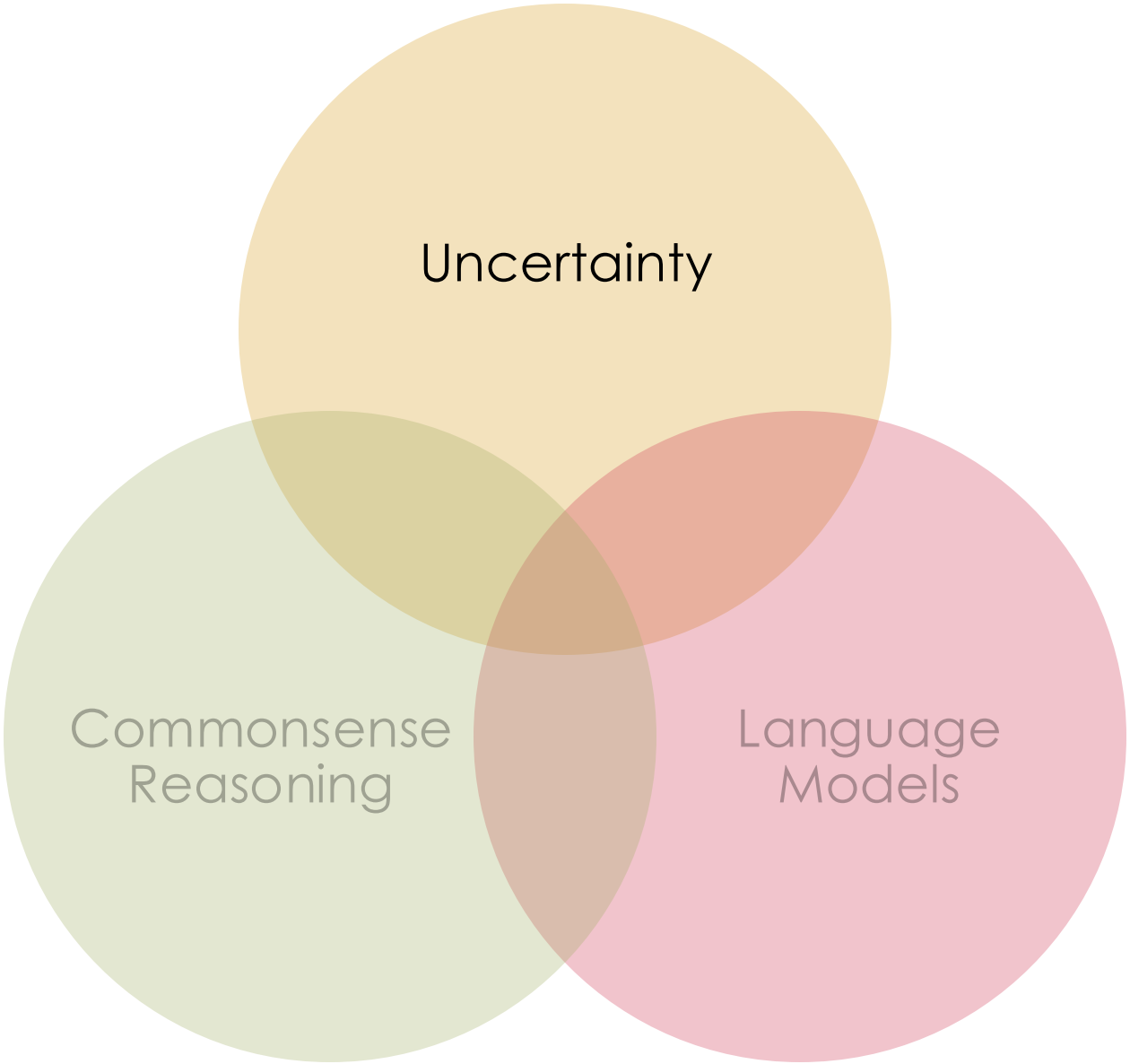
Alon Talmor[*,1,2] **Jonathan Herzig**[*,1] **Nicholas Lourie**[2] **Jonathan Berant**[1,2]




Uncertainty


Uncertainty
Commonsense
Reasoning

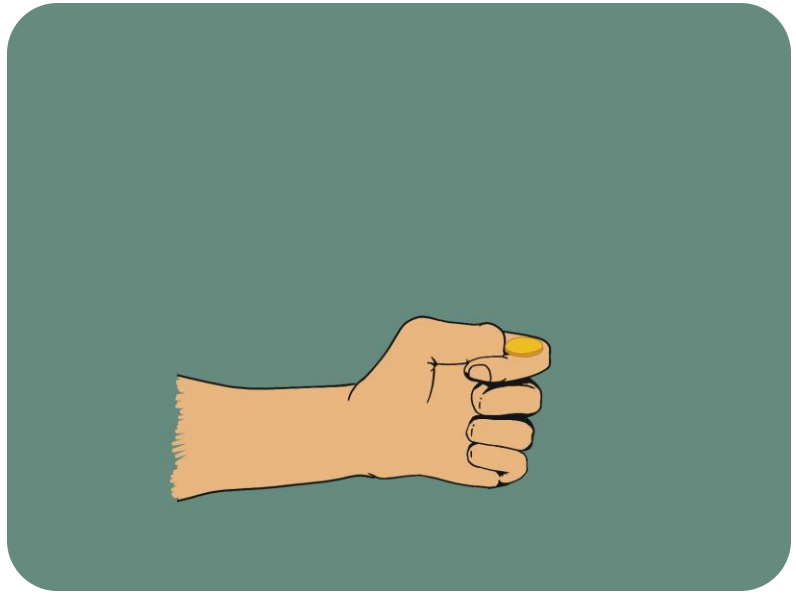

Uncertainty
Commonsense Reasoning
Language Models


Uncertainty
Commonsense Reasoning
Language Models

What is ***uncertainty***?

Let's start with an example…

What is the probability of getting heads if a fair coin is tossed?



What is ***uncertainty***?

What is the probability of getting heads if a fair coin is tossed?

How would John feel after passing a test on his second attempt?

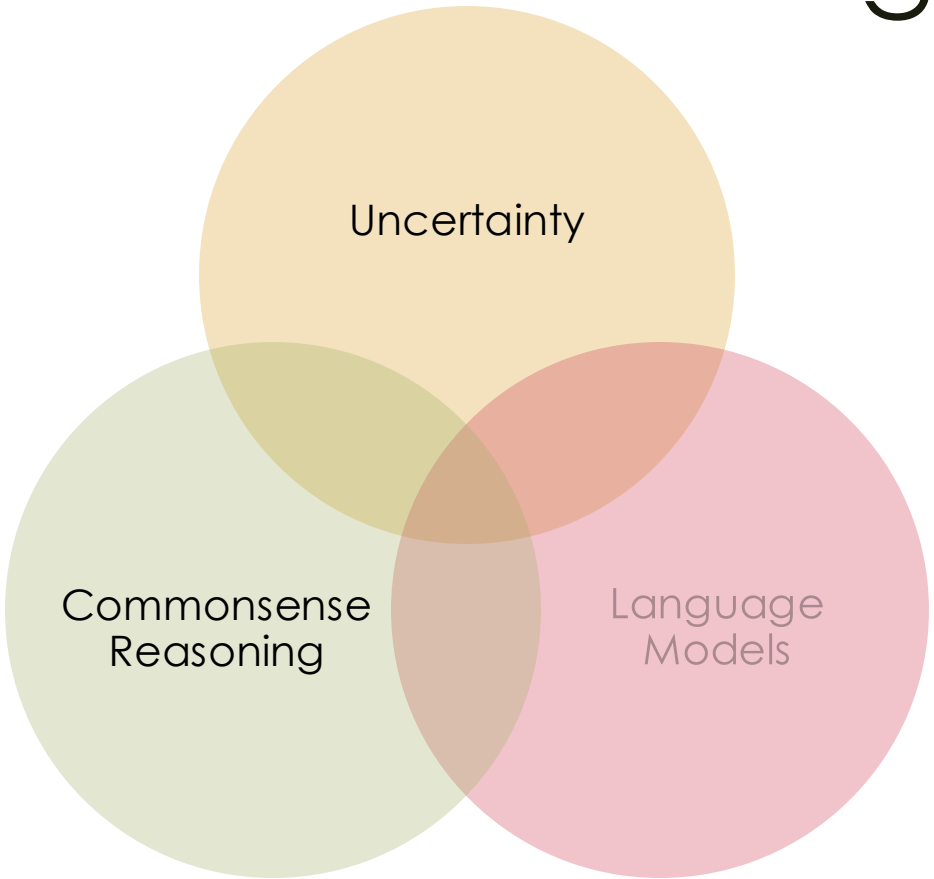

Susan went out to eat at a restaurant. She wondered whether to tip or not?



What is ***uncertainty***?

- From Wikipedia:
  - *Situations involving imperfect or unknown information.*
- Meriam Webster:
  - *Lack of sureness about someone or something.*
- APA Dictionary:
  - *Condition in which something is not accurately or precisely known.*

How does this relate to Commonsense Reasoning?



Let's take a deeper look…

Plausibly Problematic Questions in Multiple-Choice Benchmarks for Commonsense Reasoning

Shramay Palta [†] Nishant Balepur [†] Peter Rankel [†]
Sarah Wiegreffe [α] Marine Carpuat [†] Rachel Rudinger [†]

[†] University of Maryland [α] Allen Institute for Artificial Intelligence (Ai2)

EMNLP 2024

Ai2

Motivation

- Commonsense Reasoning is not free from uncertainty
- Do commonsense reasoning questions admit multiple plausible answers?
- Commonsense reasoning → soft judgements about relative *plausibility* or *likelihood* of different possible outcomes

Shramay Palta, Nishant Balepur, Peter A. Rankel, Sarah Wiegreffe, Marine Carpuat, and Rachel Rudinger. 2024. Plausibly Problematic Questions in Multiple-Choice Benchmarks for Commonsense Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*

Motivation

- Let's look at this through an example…
- What happens when a wine glass falls?

Shramay Palta, Nishant Balepur, Peter A. Rankel, Sarah Wiegreffe, Marine Carpuat, and Rachel Rudinger. 2024. Plausibly Problematic Questions in Multiple-Choice Benchmarks for Commonsense Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*

Motivation

- Let's look at this through an example…
- Very likely that it breaks…

Shramay Palta, Nishant Balepur, Peter A. Rankel, Sarah Wiegreffe, Marine Carpuat, and Rachel Rudinger. 2024. Plausibly Problematic Questions in Multiple-Choice Benchmarks for Commonsense Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*

Motivation

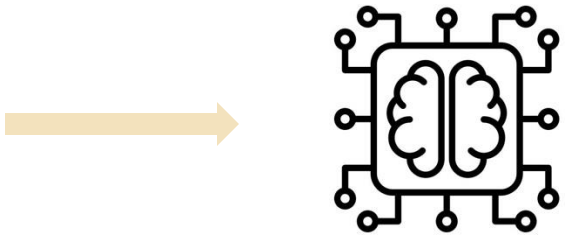
- Let's look at this through an example…
- Or… it bounces?

Shramay Palta, Nishant Balepur, Peter A. Rankel, Sarah Wiegreffe, Marine Carpuat, and Rachel Rudinger. 2024. Plausibly Problematic Questions in Multiple-Choice Benchmarks for Commonsense Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*

Motivation

- What about Multiple Choice Question (MCQ) datasets involving commonsense reasoning situations?
- Traditionally, MCQ datasets admit one correct answer.
  - *Single correct answer → easy model evaluation*

Question: What do you drink coffee in?
Choice A: Mug
Choice B: Bucket
Choice C: Kettle



Shramay Palta, Nishant Balepur, Peter A. Rankel, Sarah Wiegreffe, Marine Carpuat, and Rachel Rudinger. 2024. Plausibly Problematic Questions in Multiple-Choice Benchmarks for Commonsense Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*

So, what does it mean for a commonsense reasoning MCQ answer choice to be the "correct" answer choice?

Shramay Palta, Nishant Balepur, Peter A. Rankel, Sarah Wiegreffe, Marine Carpuat, and Rachel Rudinger. 2024. Plausibly Problematic Questions in Multiple-Choice Benchmarks for Commonsense Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*

We posit…

- Plausibility of an answer $a \rightarrow f(q, a)$
- The "correct" MCQ answer should be the one that the annotators deem to be the most plausible amongst all other options.
- What can we do with this?
  - *Let's just rate the plausibility of all answer choices for a question and choose the highest scoring option as the correct answer choice.*

Shramay Palta, Nishant Balepur, Peter A. Rankel, Sarah Wiegreffe, Marine Carpuat, and Rachel Rudinger. 2024. Plausibly Problematic Questions in Multiple-Choice Benchmarks for Commonsense Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*

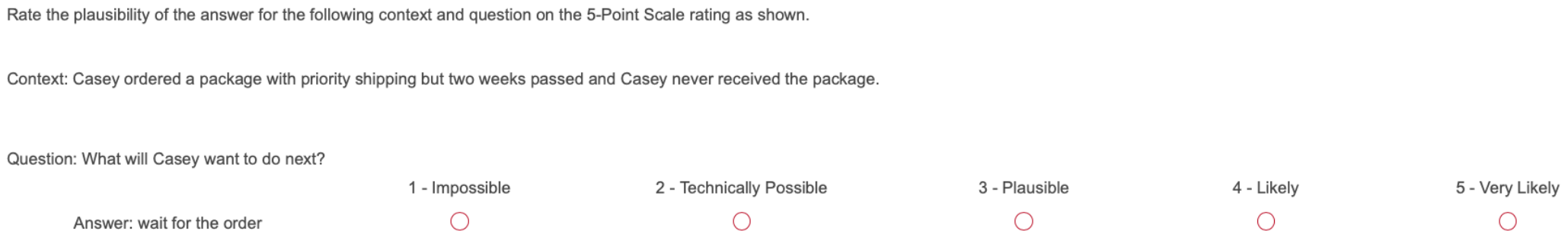
On that note

- 250 questions from two commonsense reasoning datasets: Social IQa (Sap et al. 2019) and CommonsenseQA (Talmor et al. 2019).
- Collect 5000 plausibility judgments on a 5-point Likert scale.
- Collect1530 best answer judgements for the same questions.

Shramay Palta, Nishant Balepur, Peter A. Rankel, Sarah Wiegreffe, Marine Carpuat, and Rachel Rudinger. 2024. Plausibly Problematic Questions in Multiple-Choice Benchmarks for Commonsense Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*

Plausibility Judgements

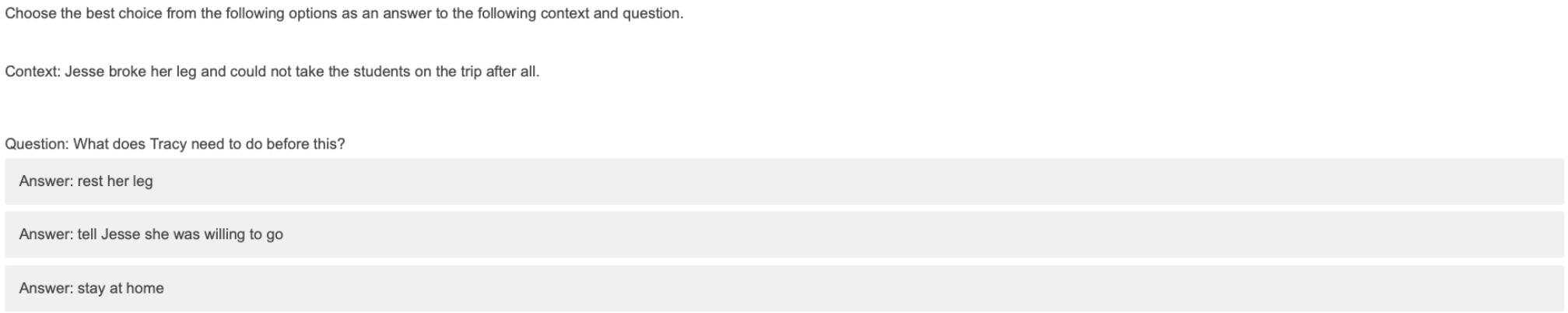
- Break down each question q with choices $a_1, a_2, \ldots a_n$ into pairs of the form (q, a_i) where $n = 3$ for Social IQa and $n = 5$ for CommonsenseQA.
- Present each (q, a) pair to annotators where they rate the plausibility of the answer choice a on a 5-point Likert scale.



Shramay Palta, Nishant Balepur, Peter A. Rankel, Sarah Wiegreffe, Marine Carpuat, and Rachel Rudinger. 2024. Plausibly Problematic Questions in Multiple-Choice Benchmarks for Commonsense Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*

Best Answer Judgements

- Present the full question with all answer choices to the annotators.
- Annotators choose the best answer as their response.
- Each MCQ item receives an initial set of 5 annotations:
  - *If no answer choices receives the majority answer vote by a margin of two or more, collect 5 additional annotations for that item.*



Shramay Palta, Nishant Balepur, Peter A. Rankel, Sarah Wiegreffe, Marine Carpuat, and Rachel Rudinger. 2024. Plausibly Problematic Questions in Multiple-Choice Benchmarks for Commonsense Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*

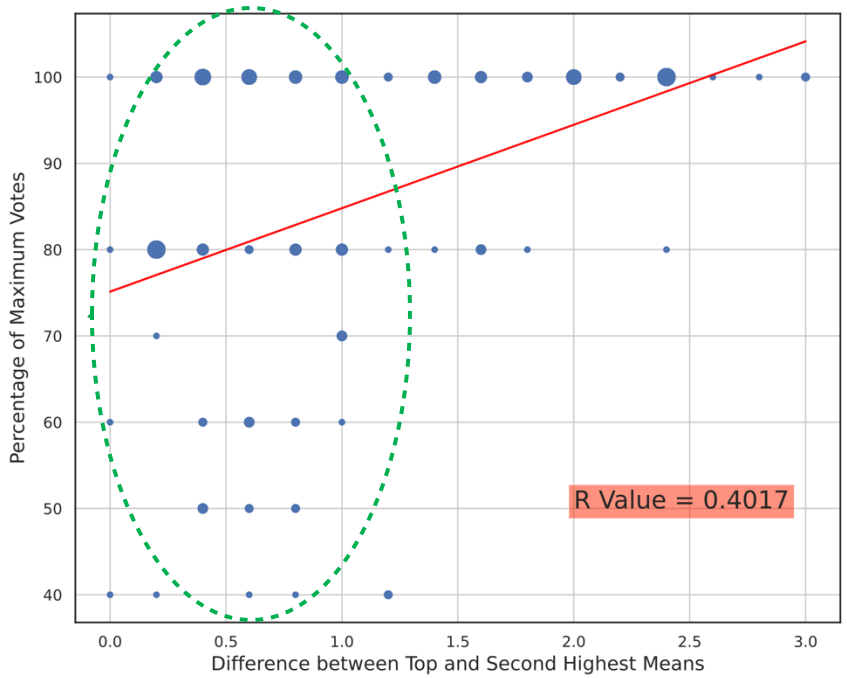
Defining "correct" answer choices

- Introduce three ways of defining a "correct" answer choices for a MCQ item:
  - $y_{dataset}$: *original gold answer label from Social IQa or CommonsenseQA*
  - $y_{plausibility}$: *answer choice with the maximum mean plausibility rating*
  - y_{full}: *majority vote answer choice from the best answer judgements*

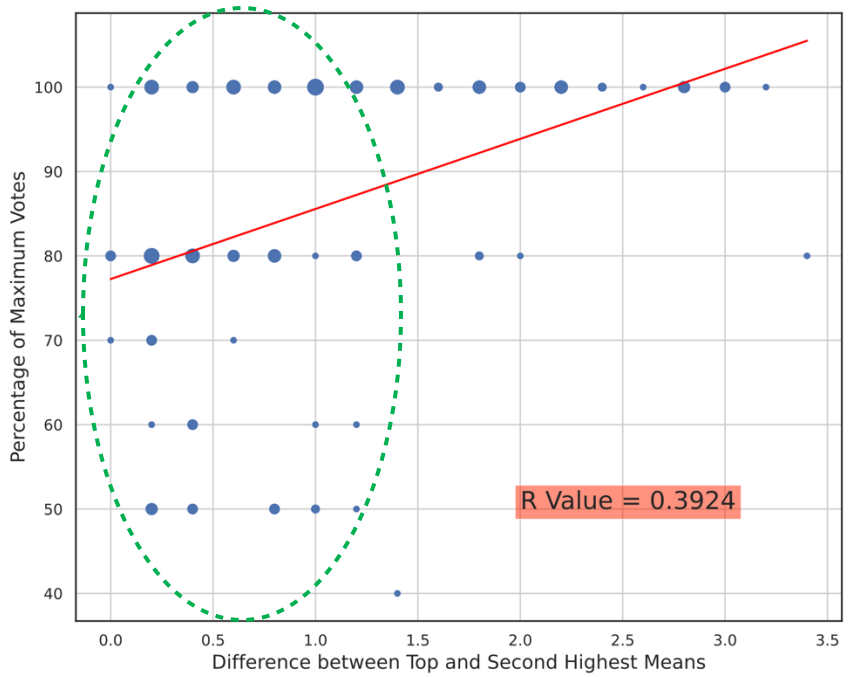
Shramay Palta, Nishant Balepur, Peter A. Rankel, Sarah Wiegreffe, Marine Carpuat, and Rachel Rudinger. 2024. Plausibly Problematic Questions in Multiple-Choice Benchmarks for Commonsense Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*

We hypothesize that when $y_{plausibility}$ is not predictive of $y_{dataset}$ and y_{full} ,it may be indicative of one or more problems with the underlying MCQ.

Shramay Palta, Nishant Balepur, Peter A. Rankel, Sarah Wiegreffe, Marine Carpuat, and Rachel Rudinger. 2024. Plausibly Problematic Questions in Multiple-Choice Benchmarks for Commonsense Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*



Social IQa



CommonsenseQA

- A small difference in the mean plausibility scores between the highest and second-highest scoring options is correlated with lower agreement on the best answer judgment setting

Shramay Palta, Nishant Balepur, Peter A. Rankel, Sarah Wiegreffe, Marine Carpuat, and Rachel Rudinger. 2024. Plausibly Problematic Questions in Multiple-Choice Benchmarks for Commonsense Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*

Building upon the initial hypothesis…

- "Plausibly Problematic" MCQs → MCQs where $y_{plausibility} \neq y_{dataset}$
  - *22.4% of the cases in both the datasets!*

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

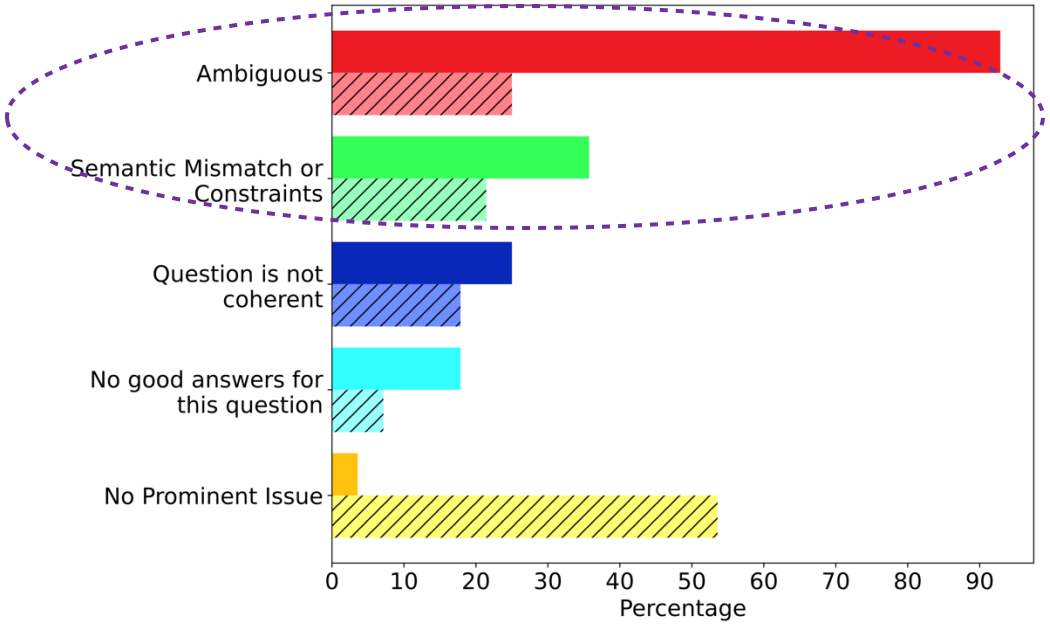
AnswerA: relieved 👴: 5, 2, 5, 5, 4 *(4.2)*
AnswerB: accomplished 👴: 4, 2, 5, 2, 5 *(3.6)*
AnswerC: proud 👴: 4, 5, 5, 5, 5 *(4.8)*

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.

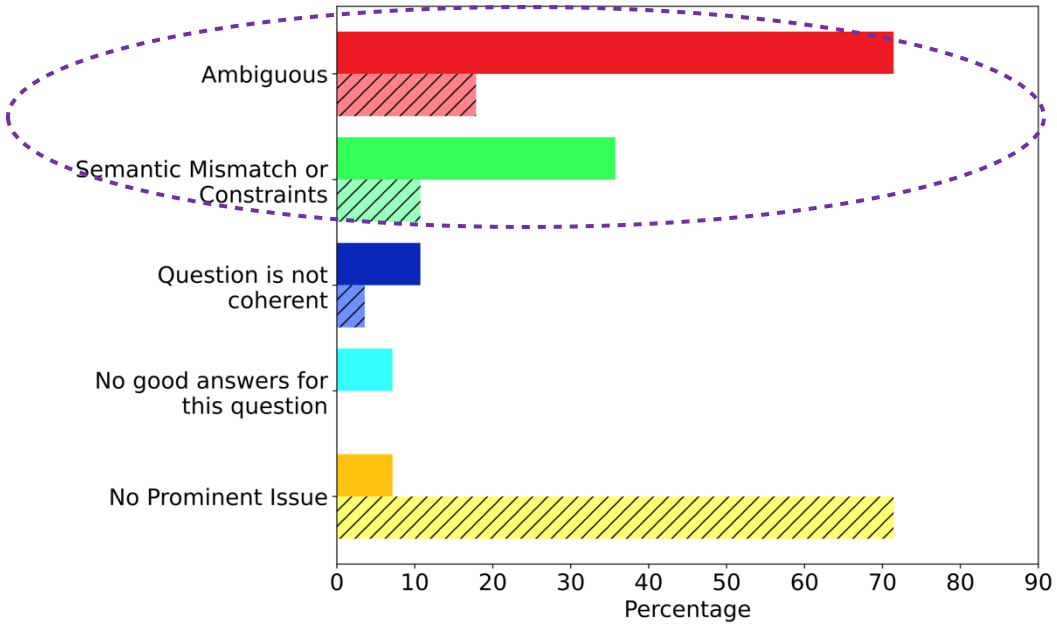
Shramay Palta, Nishant Balepur, Peter A. Rankel, Sarah Wiegreffe, Marine Carpuat, and Rachel Rudinger. 2024. Plausibly Problematic Questions in Multiple-Choice Benchmarks for Commonsense Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*

What makes a question "Plausibly Problematic"

- Manual analysis of all "plausibly problematic" questions from Social IQa and CommonsenseQA
  - *Majority of the "plausibly problematic" questions are either "Ambiguous" or have certain "Semantic Mismatch or Constraints"*



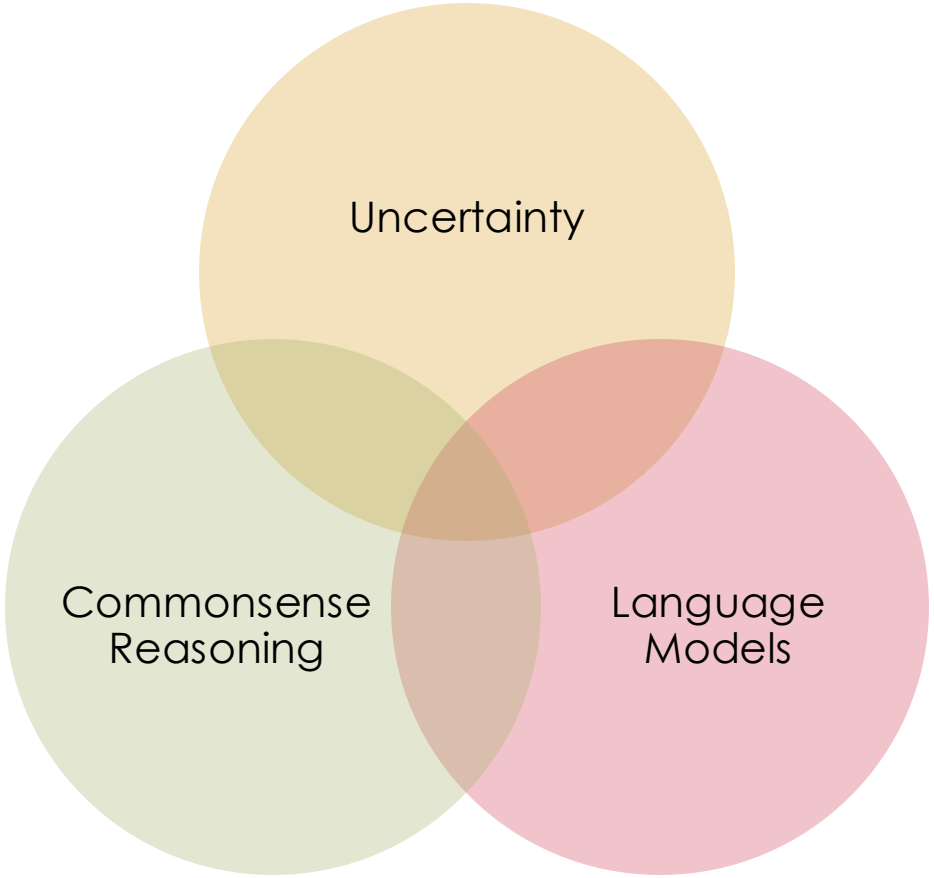
Social IQa



CommonsenseQA

Shramay Palta, Nishant Balepur, Peter A. Rankel, Sarah Wiegreffe, Marine Carpuat, and Rachel Rudinger. 2024. Plausibly Problematic Questions in Multiple-Choice Benchmarks for Commonsense Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*

Impact of Uncertainty in Commonsense Reasoning



Testing LLMs on this setup

- Multiple state-of-the-art LLMs:
  - *GPT-4, LLaMA 2, Mistral, Yi*

- 10 in-context examples for prompting
  - *Examples with lowest variance in human ratings for each Likert scale rating*

Shramay Palta, Nishant Balepur, Peter A. Rankel, Sarah Wiegreffe, Marine Carpuat, and Rachel Rudinger. 2024. Plausibly Problematic Questions in Multiple-Choice Benchmarks for Commonsense Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*

Implications for LLM Evaluation

Agent	SIQA			CSQA		
	Prob	Non	All	Prob	Non	All
LLaMA-2 7B	53.8	67.4	64.3	55.6	67.0	64.3
LLaMA-2 13B	42.3	75.3	67.8	55.6	77.3	72.2
LLaMA-2 70B	57.7	87.6	80.9	66.7	85.2	80.9
Mistral 7B	38.5	80.9	71.3	59.3	76.1	72.2
Mixtral 7x8B	53.8	86.5	79.1	66.7	87.5	82.6
Yi 6B	50.0	84.3	76.5	63.0	84.1	79.1
Yi 9B	73.1	91.0	87.0	74.1	85.2	82.6
Yi 34B	61.5	94.4	87.0	70.4	90.9	86.1
GPT-4	53.8	89.9	81.7	59.3	92.0	84.3
Average LLM	53.8	84.1	77.3	63.4	82.8	78.3
Human	71.2	94.4	89.1	70.4	92.6	87.4

- LLMs have lower accuracy on the set of "plausibly problematic" questions!
- Performance drop in the problematic set is much larger for LLMs!

Shramay Palta, Nishant Balepur, Peter A. Rankel, Sarah Wiegreffe, Marine Carpuat, and Rachel Rudinger. 2024. Plausibly Problematic Questions in Multiple-Choice Benchmarks for Commonsense Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*

Takeaways

- Plausibility judgements are a reliable tool for identifying "plausibly problematic" MCQ test items.
- Individual plausibility ratings reveal several issues with MCQ items that are more prevalent in the problematic subset of questions.
- LLMs and humans both perform poorly on this subset.

Shramay Palta, Nishant Balepur, Peter A. Rankel, Sarah Wiegreffe, Marine Carpuat, and Rachel Rudinger. 2024. Plausibly Problematic Questions in Multiple-Choice Benchmarks for Commonsense Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*

Can something increase or decrease this plausibility?

- So far, plausibility of a choice $a \rightarrow f(q, a)$

- Is there any additional stimulus that can influence these plausibility ratings?

$$f(q, a) \rightarrow f(q, a, ?)$$

Let's find out more…

Everything is Plausible: Investigating the Impact of LLM Rationales on Human Notions of Plausibility

Shramay Palta † Peter Rankel †
Sarah Wiegreffe[α] Rachel Rudinger †

† University of Maryland α Allen Institute for Artificial Intelligence (Ai2)

Motivation

- Let's go back to our previous example…
- What happens when a wine glass falls?

Shramay Palta, Peter Rankel, Sarah Wiegreffe and Rachel Rudinger. 2025. Everything is Plausible: Investigating the Impact of LLM Rationales of Human Notions of Plausibility. Under Review

Motivation

- Glass doesn't break because it's sturdy or it do fall far.

Shramay Palta, Peter Rankel, Sarah Wiegreffe and Rachel Rudinger. 2025. Everything is Plausible: Investigating the Impact of LLM Rationales of Human Notions of Plausibility. Under Review

Motivation

- Glass bounces because it lands on a trampoline or a rubber mat.

Shramay Palta, Peter Rankel, Sarah Wiegreffe and Rachel Rudinger. 2025. Everything is Plausible: Investigating the Impact of LLM Rationales of Human Notions of Plausibility. Under Review

Motivation

- Circumstances matter!
- A relatively implausible distractor could be true.
  - *Glass bounces because it landed on a trampoline or a rubber mat*
- A highly plausible outcome could become less likely.
  - *Glass doesn't break because it's sturdy*

Shramay Palta, Peter Rankel, Sarah Wiegreffe and Rachel Rudinger. 2025. Everything is Plausible: Investigating the Impact of LLM Rationales of Human Notions of Plausibility. Under Review

An answer choice can be subject to arguments for or against its plausibility

Shramay Palta, Peter Rankel, Sarah Wiegreffe and Rachel Rudinger. 2025. Everything is Plausible: Investigating the Impact of LLM Rationales of Human Notions of Plausibility. Under Review

Motivation

- These (im)plausibility arguments do not add any new evidence!
- Highlight possible circumstances, which *if true*, would impact an answer's plausibility

Shramay Palta, Peter Rankel, Sarah Wiegreffe and Rachel Rudinger. 2025. Everything is Plausible: Investigating the Impact of LLM Rationales of Human Notions of Plausibility. Under Review

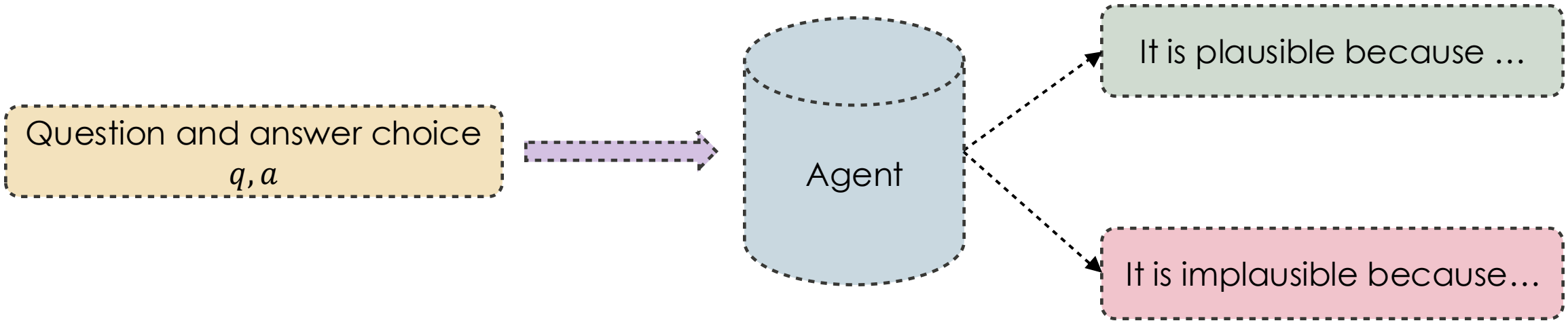
Let's add that stimulus

- Plausibility of an answer choice $a \rightarrow f(q, a, r)$
  - $r \rightarrow$ *argument in favor or against the plausibility of* a

- How do we generate these (im)plausibility arguments r?

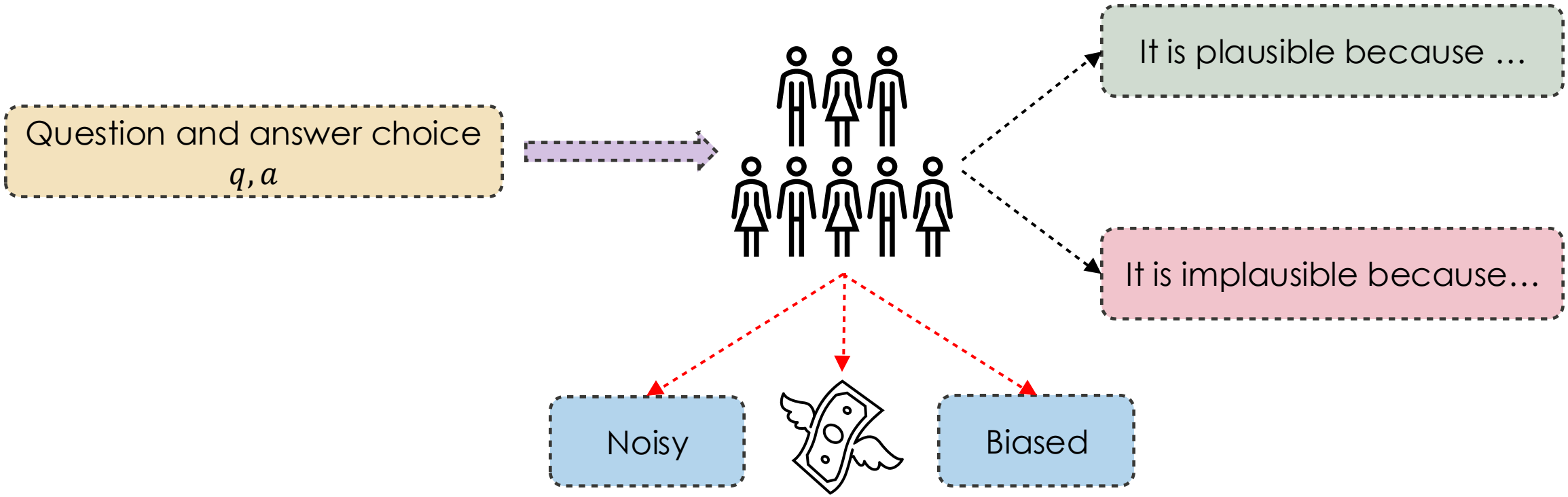
Shramay Palta, Peter Rankel, Sarah Wiegreffe and Rachel Rudinger. 2025. Everything is Plausible: Investigating the Impact of LLM Rationales of Human Notions of Plausibility. Under Review

Generating rationales r



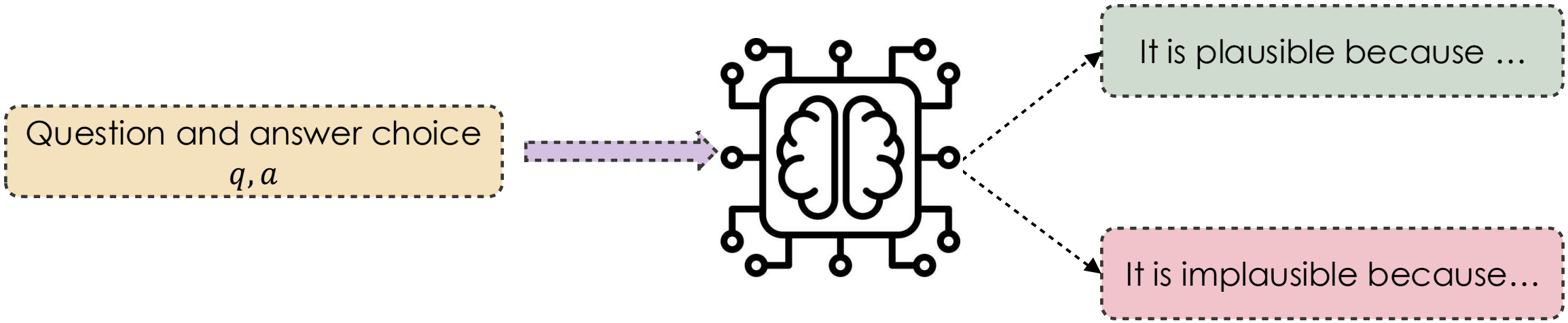
Shramay Palta, Peter Rankel, Sarah Wiegreffe and Rachel Rudinger. 2025. Everything is Plausible: Investigating the Impact of LLM Rationales of Human Notions of Plausibility. Under Review

Generating rationales r



Shramay Palta, Peter Rankel, Sarah Wiegreffe and Rachel Rudinger. 2025. Everything is Plausible: Investigating the Impact of LLM Rationales of Human Notions of Plausibility. Under Review

Generating rationales r



Shramay Palta, Peter Rankel, Sarah Wiegreffe and Rachel Rudinger. 2025. Everything is Plausible: Investigating the Impact of LLM Rationales of Human Notions of Plausibility. Under Review

Which LLM though?

- No shortage of modern-day Large Language Models to choose from.

- How do we decide which LLM to show to annotators?

Shramay Palta, Peter Rankel, Sarah Wiegreffe and Rachel Rudinger. 2025. Everything is Plausible: Investigating the Impact of LLM Rationales of Human Notions of Plausibility. Under Review

Preference Study

- Prompt 4 models to generate two types of rationales r:
  - r_{pro} : *rationale in favor of plausibility of the answer choice* a
  - r_{con} : *rationale in favor of implausibility of the answer choice* a
- Present r_{pro} and r_{con} and their corresponding (q, a) pairs to 4 annotators.
- Models: GPT-4o, GPT-4o-mini, LLaMa-3.1-(8B, 70B)-Instruct
- GPT-4o → most votes → LLM of choice!

Shramay Palta, Peter Rankel, Sarah Wiegreffe and Rachel Rudinger. 2025. Everything is Plausible: Investigating the Impact of LLM Rationales of Human Notions of Plausibility. Under Review

Human Plausibility Ratings

- Sample 50 'non-problematic' questions each from both Social IQa and CommonsenseQA from the previous study.

- For each question q, consider two answer choices:
  - $a_{gold\ label}$: *original dataset gold label*
  - $a_{distractor}$: *randomly sampled distractor*

- Total 100 (q, a) pairs per dataset, where $a \in [a_{gold\ label}, a_{distractor}]$

Shramay Palta, Peter Rankel, Sarah Wiegreffe and Rachel Rudinger. 2025. Everything is Plausible: Investigating the Impact of LLM Rationales of Human Notions of Plausibility. Under Review

Human Plausibility Ratings

- 4 types of plausibility ratings:
  - *No Rationale Plausibility Rating*
  - *PRO Rationale Plausibility Rating*
  - *CON Rationale Plausibility Rating*
  - *PRO+CON Rationale Plausibility Rating*

- 5 annotators per (q, a, r) pair where $r \in [r_{pro}, r_{con}, r_{pro\ +\ con}]$

- 3000 human judgments collected

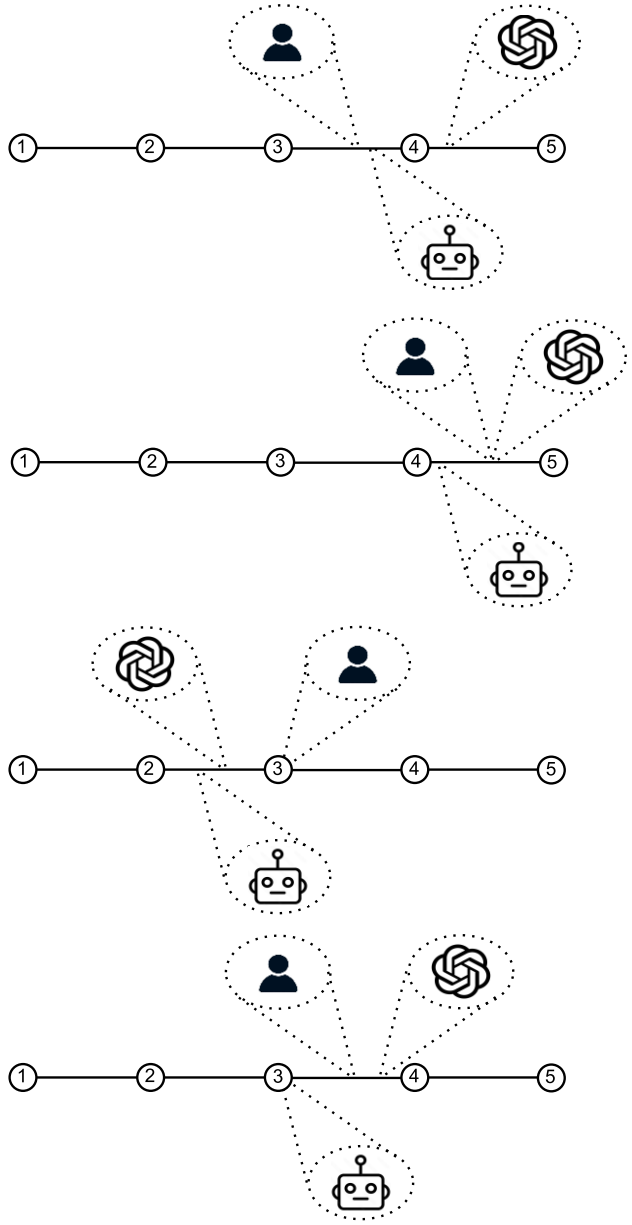
Shramay Palta, Peter Rankel, Sarah Wiegreffe and Rachel Rudinger. 2025. Everything is Plausible: Investigating the Impact of LLM Rationales of Human Notions of Plausibility. Under Review

Question: If a car-less person want to listen to talk radio in private, where might they listen to it?
Choice: bedroom (gold label)

Pro Rationale: A bedroom is a private space where a car-less person can use a radio, smartphone, or other devices to tune into talk radio without external disturbances, ensuring an undisturbed listening experience. Additionally, bedrooms are typically associated with comfort and quiet, reinforcing the ability to focus on the content.

Con Rationale: The bedroom is an implausible choice because it is a communal space shared with others in many households, which makes it difficult to ensure full privacy for listening to talk radio. Additionally, a person might not have access to a radio or leisure time in their own bedroom if they share living accommodations.

Pro Rationale: A bedroom is a private space ...
Con Rationale: The bedroom is an implausible choice ...



Shramay Palta, Peter Rankel, Sarah Wiegreffe and Rachel Rudinger. 2025. Everything is Plausible: Investigating the Impact of LLM Rationales of Human Notions of Plausibility. Under Review

Impact on Human Ratings

Dataset	Pro Rationale			Con Rationale			Pro+Con Rationales		
	Increase	Decrease	No Change	Increase	Decrease	No Change	Increase	Decrease	No Change
SIQA	28%	22%	50%	2%	69%	29%	11%	44%	45%
CQA	30%	27%	43%	9%	33%	58%	13%	44%	43%

- r_{pro} leads to an increase in the mean plausibility ratings.
- r_{con} leads to a decrease in the mean plausibility ratings.
- $r_{pro\ +\ con}$ leads to either a decrease or no change in the mean plausibility ratings.

- Chi-squared tests of homogeneity highlight high statistical significance in almost all cases.

Shramay Palta, Peter Rankel, Sarah Wiegreffe and Rachel Rudinger. 2025. Everything is Plausible: Investigating the Impact of LLM Rationales of Human Notions of Plausibility. Under Review

How does this affect LLM ratings?

- Two families of models
  - *Issues of self preference!*

OpenAI Models	Non-OpenAI Models
GPT-(3.5, 4, 4-turbo, 4o, 4o-mini, 4.5-preview, o1, o3-mini)	LLaMa-3.1-Instruct, LLaMa-3.2-Instruct, LLaMa-3.3-Instruct, Mistral Instruct, Yi-1.5-Chat, DeepSeek-R1

- Zero-shot prompts
- 13,600 LLM judgments collected

Shramay Palta, Peter Rankel, Sarah Wiegreffe and Rachel Rudinger. 2025. Everything is Plausible: Investigating the Impact of LLM Rationales of Human Notions of Plausibility. Under Review

Impact on LLM Ratings

Dataset	Pro Rationale			Con Rationale			Pro+Con Rationales		
	Increase	Decrease	No Change	Increase	Decrease	No Change	Increase	Decrease	No Change
SIQA	40%	2%	58%	0%	84%	16%	12%	43%	45%
CQA	62%	4%	34%	4%	74%	22%	21%	38%	41%

OpenAI Models

Dataset	Pro Rationale			Con Rationale			Pro+Con Rationales		
	Increase	Decrease	No Change	Increase	Decrease	No Change	Increase	Decrease	No Change
SIQA	61%	1%	38%	5%	66%	29%	20%	24%	56%
CQA	53%	5%	42%	4%	65%	31%	21%	32%	47%

Non-OpenAI Models

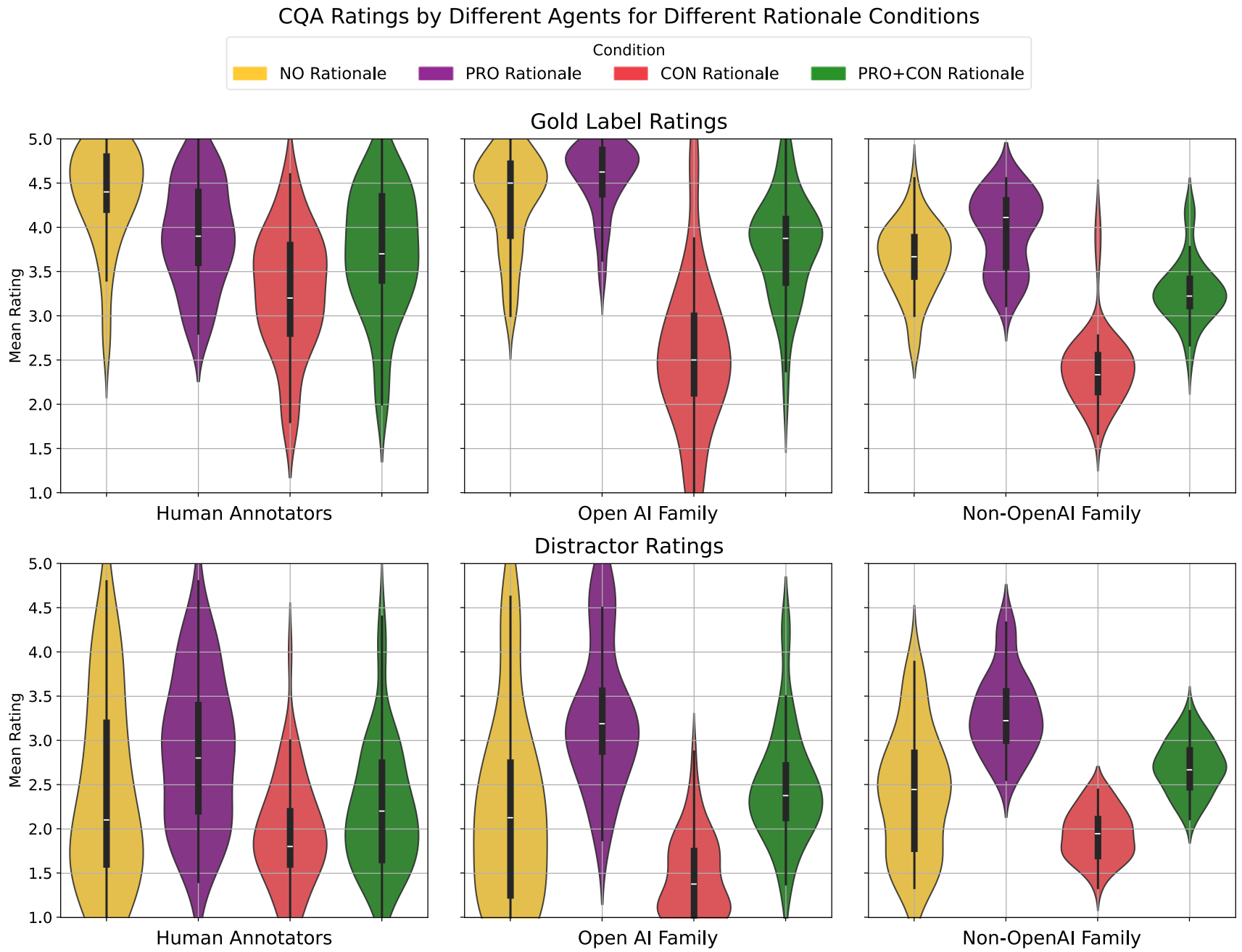
Shramay Palta, Peter Rankel, Sarah Wiegreffe and Rachel Rudinger. 2025. Everything is Plausible: Investigating the Impact of LLM Rationales of Human Notions of Plausibility. Under Review

Is that it?

- Answer choice $a \in [a_{gold\ label}, a_{distractor}]$
- What about individual effects on these choices?

Shramay Palta, Peter Rankel, Sarah Wiegreffe and Rachel Rudinger. 2025. Everything is Plausible: Investigating the Impact of LLM Rationales of Human Notions of Plausibility. Under Review

Shramay Palta, Peter Rankel, Sarah Wiegreffe and Rachel Rudinger. 2025. Everything is Plausible: Investigating the Impact of LLM Rationales of Human Notions of Plausibility. Under Review



Shramay Palta, Peter Rankel, Sarah Wiegreffe and Rachel Rudinger. 2025. Everything is Plausible: Investigating the Impact of LLM Rationales of Human Notions of Plausibility. Under Review

What do these results imply?

- Are humans reliable judges of commonsense reasoning situations?
- Shifts due to better arguments or framing/cognitive biases?
- LLMs are creative!

Shramay Palta, Peter Rankel, Sarah Wiegreffe and Rachel Rudinger. 2025. Everything is Plausible: Investigating the Impact of LLM Rationales of Human Notions of Plausibility. Under Review

Summary

- Both human and LLM plausibility ratings shift with inclusion of different types of rationales.
- Humans and LLMs respond differently to this stimulus.
- Can LLMs help humans?

Shramay Palta, Peter Rankel, Sarah Wiegreffe and Rachel Rudinger. 2025. Everything is Plausible: Investigating the Impact of LLM Rationales of Human Notions of Plausibility. Under Review

Can this uncertainty lead to any biases or stereotypes?

FORK: A Bite-Sized Test Set for Probing Culinary Cultural Biases in Commonsense Reasoning Models

Shramay Palta Rachel Rudinger

University of Maryland

Motivation

- Some types of knowledge are sufficiently generic to be shared by most people around the world.
  - *Commonsense Knowledge like objects fall down when they are dropped.*
- But the catch is:
  - *Who counts as "most" people?*

Shramay Palta and Rachel Rudinger. 2023. FORK: A Bite-Sized Test Set for Probing Culinary Cultural Biases in Commonsense Reasoning Models. In *Findings of the Association for Computational Linguistics: ACL 2023*

A Classic Example

- Restaurant Script (Schank and Abelson, 1975)
  - *LEAVE TIP event.*

OR

Shramay Palta and Rachel Rudinger. 2023. FORK: A Bite-Sized Test Set for Probing Culinary Cultural Biases in Commonsense Reasoning Models. In *Findings of the Association for Computational Linguistics: ACL 2023*

Motivation

- Cultural context can be one of the several reasons why a commonsense reasoning situation seems to have uncertainty.
- Several past works on commonsense knowledge acquisition rely on crowdsourcing, corpus statistics and language modeling.
- *What about the implicit cultural perspectives of corpus texts, crowd workers or AI researchers?*

Shramay Palta and Rachel Rudinger. 2023. FORK: A Bite-Sized Test Set for Probing Culinary Cultural Biases in Commonsense Reasoning Models. In *Findings of the Association for Computational Linguistics: ACL 2023*

What is 'culture'?

- Like commonsense knowledge, culture is vast…
- From Wikipedia:
  - *culture "encompasses the social behavior and norms found in human societies, as well as knowledge, beliefs, arts, laws, customs, capabilities and habits of the individuals in these groups."*
- From the social sciences: culture encompasses both material and non-material aspects, such as beliefs and linguistic practices (Kendall 2015).

Can we determine the cultural contingency of commonsense reasoning models?

Shramay Palta and Rachel Rudinger. 2023. FORK: A Bite-Sized Test Set for Probing Culinary Cultural Biases in Commonsense Reasoning Models. In *Findings of the Association for Computational Linguistics: ACL 2023*

We introduce…

- **FORK**: **F**ood **OR**iented cultural commonsense **K**nowledge
  - *Manually curated test set of 184 CommonsenseQA-style questions.*
  - *Food and culinary cultures and practices of US and several other countries.*
  - *Themes spanning across restaurant tipping, eating utensils and other culinary customs.*

OR

Shramay Palta and Rachel Rudinger. 2023. FORK: A Bite-Sized Test Set for Probing Culinary Cultural Biases in Commonsense Reasoning Models. In *Findings of the Association for Computational Linguistics: ACL 2023*

FORK

- **FORK**: **F**ood **OR**iented cultural commonsense **K**nowledge
  - *Each question has 2 options, out of which only one is correct*
  - *One US option and one Non-US option*
  - *Three types of questions:*
    - Underspecified
    - Implicit
    - Explicit

Shramay Palta and Rachel Rudinger. 2023. FORK: A Bite-Sized Test Set for Probing Culinary Cultural Biases in Commonsense Reasoning Models. In *Findings of the Association for Computational Linguistics: ACL 2023*

FORK

Q1: While eating, when does one drink soup? [Underspecified]
Q2: While eating, when does one drink Cantonese seafood soup? [Implicit]
Q3: While eating in China/the United States, when does one drink soup? [Explicit]

A1: Before the main dish. [United States]
A2: After the main dish. [China]

An example question from FORK

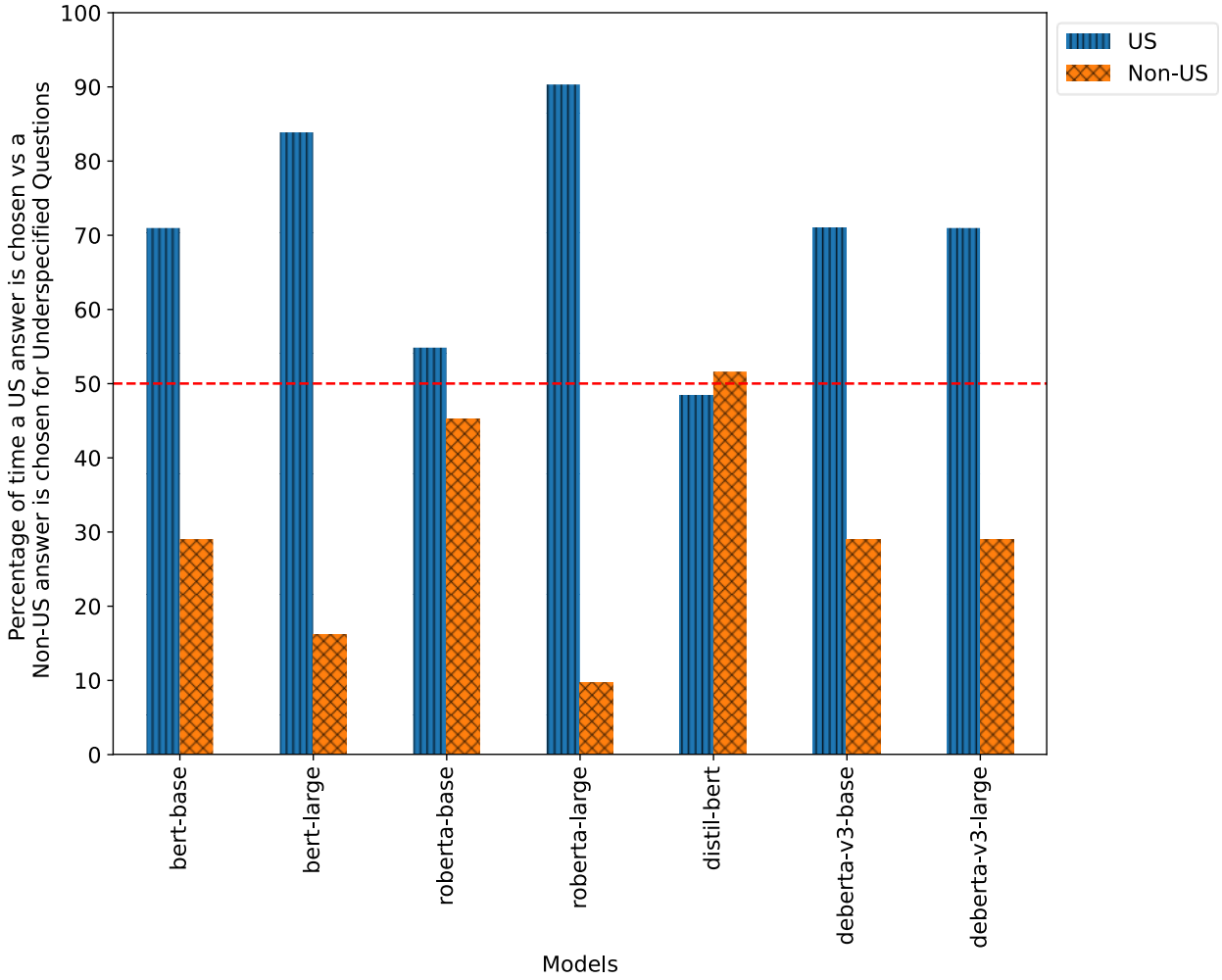
Shramay Palta and Rachel Rudinger. 2023. FORK: A Bite-Sized Test Set for Probing Culinary Cultural Biases in Commonsense Reasoning Models. In *Findings of the Association for Computational Linguistics: ACL 2023*

Evaluation Strategy

- Test 7 models from the BERT family finetuned on CommonsenseQA.
- Underspecified Questions: number of times a US answer is chosen over a Non-US answer.
- Implicit and Explicit Questions: percentage accuracy for US vs Non-US answers.

Shramay Palta and Rachel Rudinger. 2023. FORK: A Bite-Sized Test Set for Probing Culinary Cultural Biases in Commonsense Reasoning Models. In *Findings of the Association for Computational Linguistics: ACL 2023*

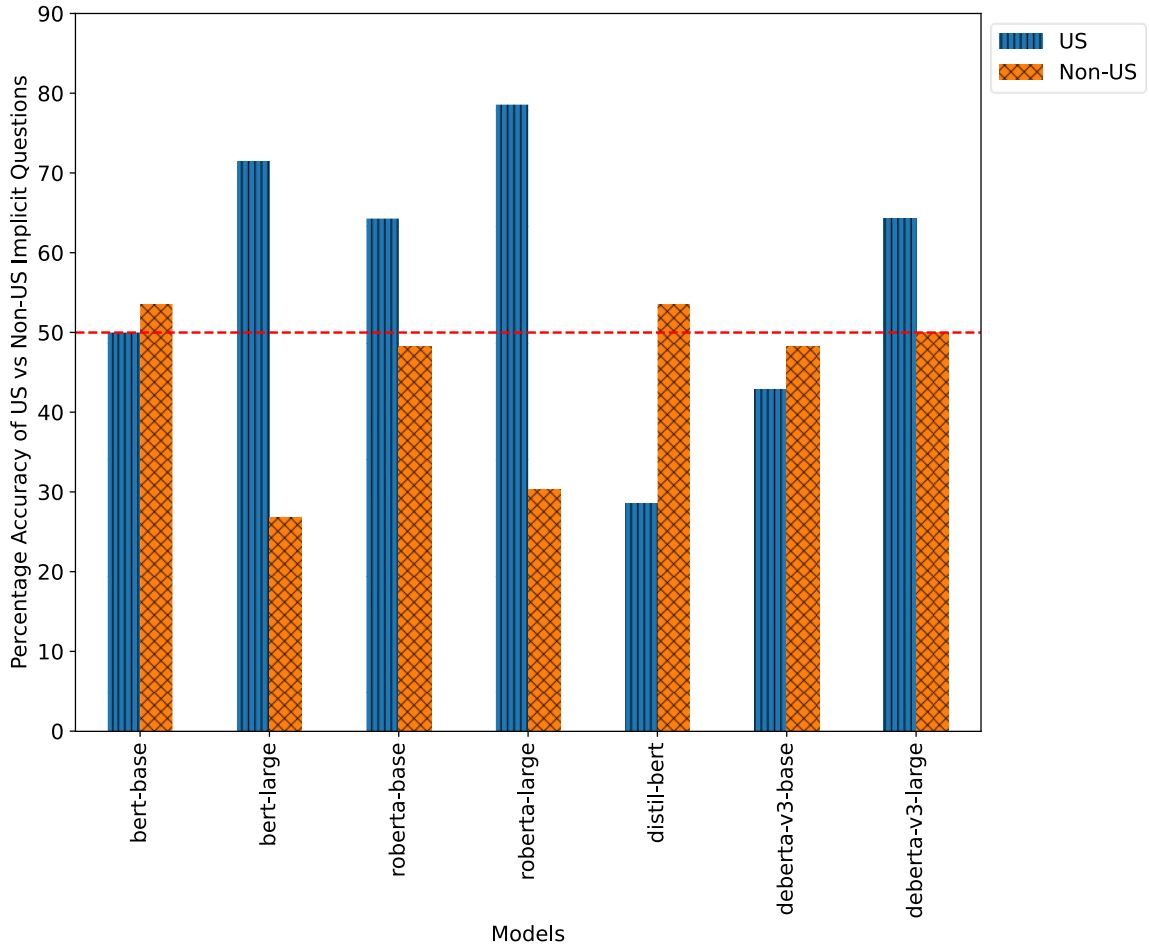
Key Results



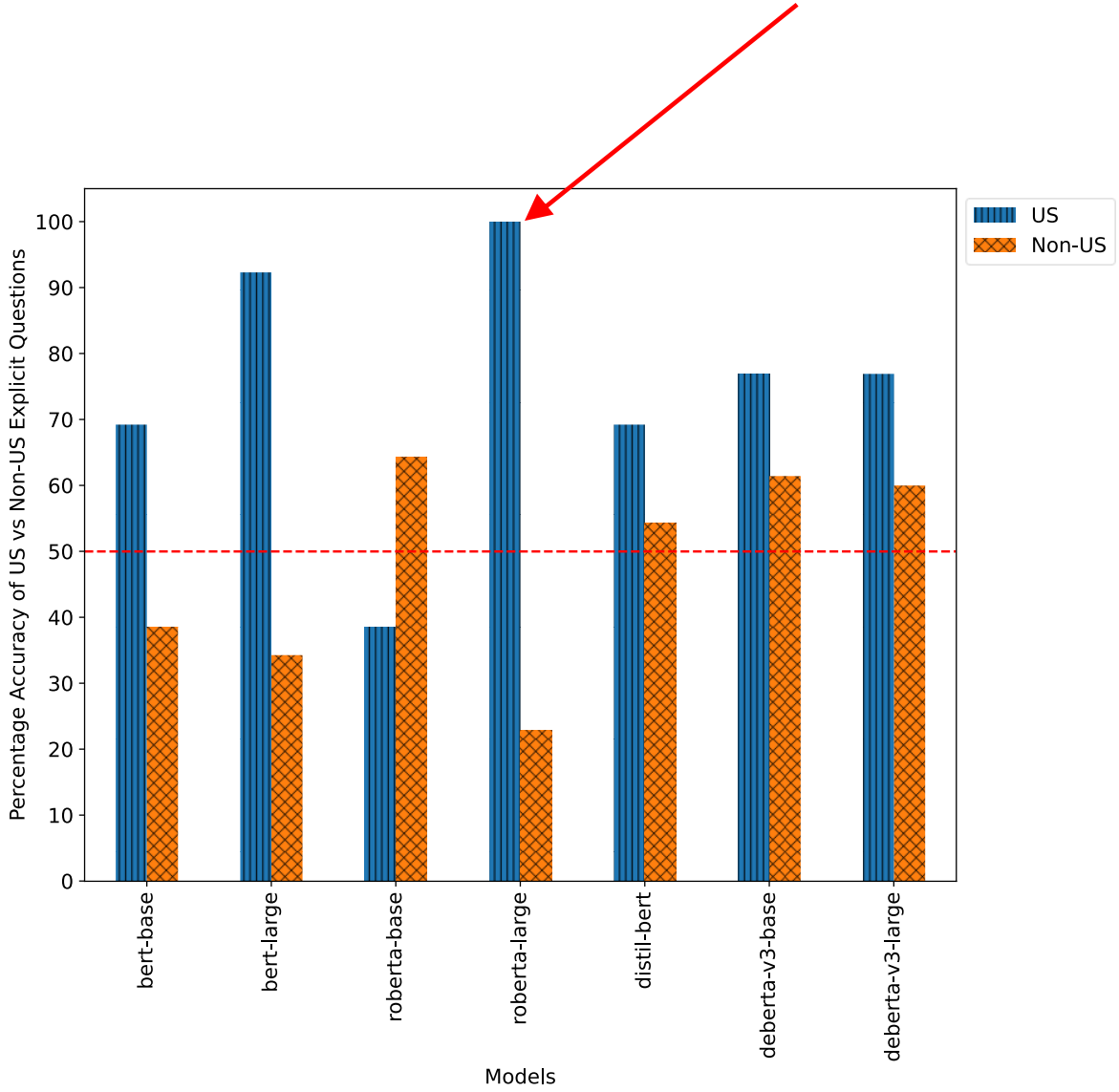
Performance on Underspecified Questions

Shramay Palta and Rachel Rudinger. 2023. FORK: A Bite-Sized Test Set for Probing Culinary Cultural Biases in Commonsense Reasoning Models. In *Findings of the Association for Computational Linguistics: ACL 2023*

Key Results



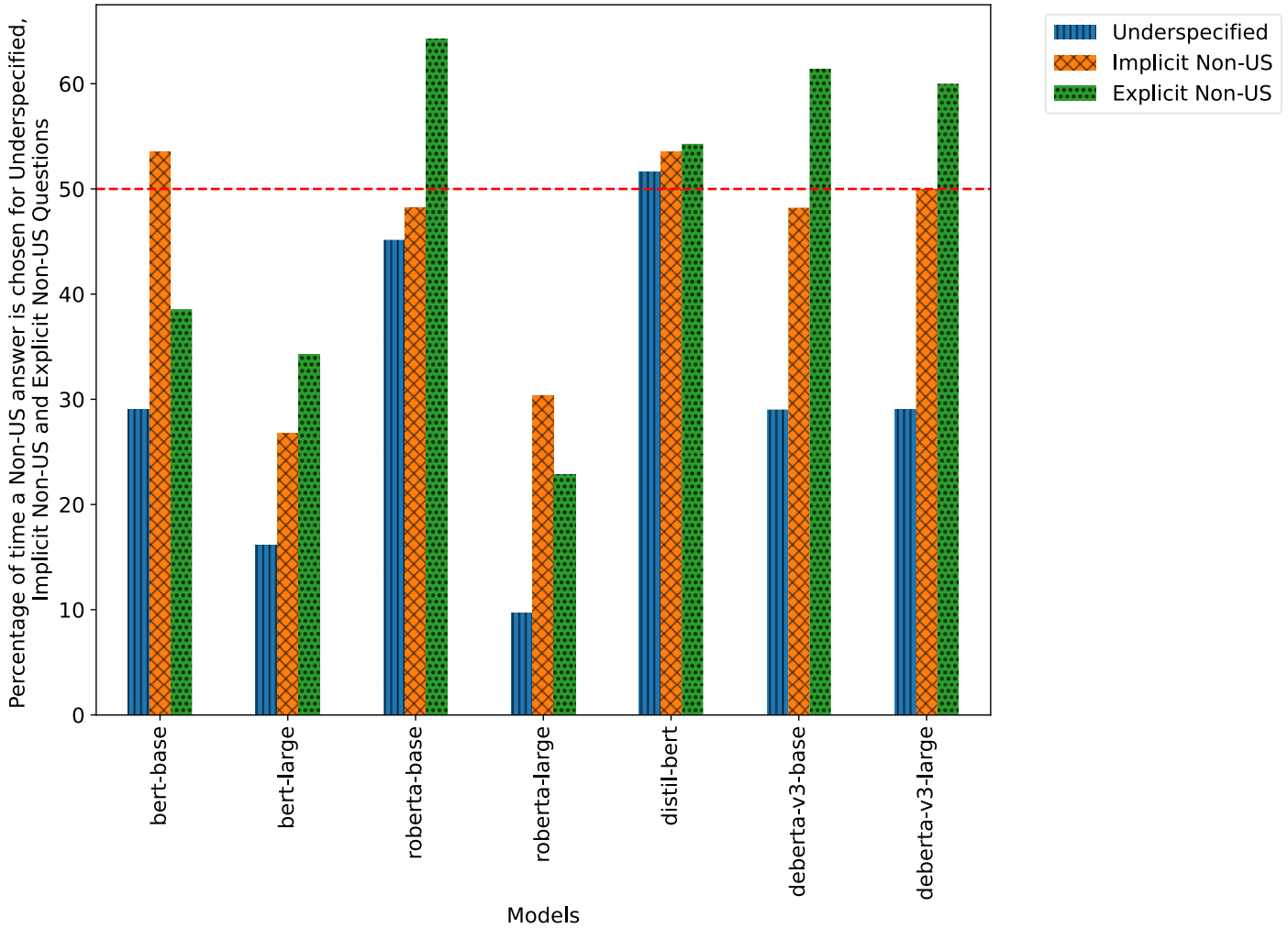
Performance on Implicit Questions



Performance on Explicit Questions

Shramay Palta and Rachel Rudinger. 2023. FORK: A Bite-Sized Test Set for Probing Culinary Cultural Biases in Commonsense Reasoning Models. In *Findings of the Association for Computational Linguistics: ACL 2023*

Key Results



Percentage of times a Non-Us answer is chosen for different types of questions.

Shramay Palta and Rachel Rudinger. 2023. FORK: A Bite-Sized Test Set for Probing Culinary Cultural Biases in Commonsense Reasoning Models. In *Findings of the Association for Computational Linguistics: ACL 2023*

Takeaways

- Uncertainty in commonsense reasoning situations can lead to biases and stereotypes in predictions by Language Models.
- Models finetuned for commonsense reasoning can make cultural assumptions.
- FORK: a new test set for evaluating cultural contingency of models.
- FORK helps demonstrate systematic cultural biases favoring US over Non-US cultures.

Shramay Palta and Rachel Rudinger. 2023. FORK: A Bite-Sized Test Set for Probing Culinary Cultural Biases in Commonsense Reasoning Models. In *Findings of the Association for Computational Linguistics: ACL 2023*

Huge Thanks to all my Collaborators!

My amazing advisor :)

UNIVERSITY OF MARYLAND 18 56

UNIVERSITY OF MARYLAND 18 56
INSTITUTE FOR ADVANCED COMPUTER STUDIES





Ai2

Microsoft Research