# Plausibly Problematic Questions in Multiple-Choice Benchmarks for Commonsense Reasoning

**Shramay Palta**♠ **Nishant Balepur**♠ **Peter Rankel**♠
**Sarah Wiegreffe**♢ **Marine Carpuat**♠ **Rachel Rudinger**♠

♠University of Maryland, College Park ♢Allen Institute for AI (AI2)
{spalta,nbalepur,par,marine,rudinger}@umd.edu
wiegreffesarah@gmail.com

## Abstract

Questions involving commonsense reasoning about everyday situations often admit many *possible* or *plausible* answers. In contrast, multiple-choice question (MCQ) benchmarks for commonsense reasoning require a hard selection of a single correct answer, which, in principle, should represent the *most* plausible answer choice. On 250 MCQ items sampled from two commonsense reasoning benchmarks, we collect 5,000 independent plausibility judgments on answer choices. We find that for over 20% of the sampled MCQs, the answer choice rated most plausible does not match the benchmark gold answers; upon manual inspection, we confirm that this subset exhibits higher rates of problems like ambiguity or semantic mismatch between question and answer choices. Experiments with LLMs reveal low accuracy and high variation in performance on the subset, suggesting our plausibility criterion may be helpful in identifying more reliable benchmark items for commonsense evaluation.[1]

## 1 Introduction

Commonsense reasoning about everyday situations involves soft judgments about the relative *plausibility* or *likelihood* of different possible outcomes. If a wine glass falls, a *very likely* outcome is that it breaks, but another *technically possible* outcome is that it bounces (e.g., because it lands on a trampoline). Datasets like the Choice of Plausible Alternatives (COPA; Roemmele et al., 2011) or Ordinal Common-sense Inference (Zhang et al., 2017) highlight this graded nature of commonsense reasoning. Many recently developed benchmark datasets for commonsense reasoning formulate problems as multiple choice questions (MCQs): PIQA (Bisk et al., 2020), Social IQa (Sap et al., 2019), CommonsenseQA (Talmor et al., 2019), among others.

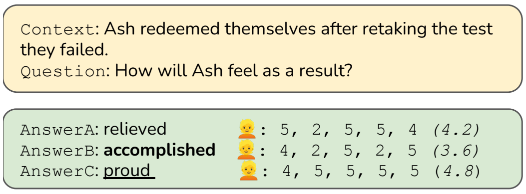



Figure 1: An example question from Social IQa where the highest plausibility answer choice is not the gold label. The numbers indicate the plausibility ratings given by 5 human annotators to each option on a 1-5 scale and the gold label is highlighted in bold. Numbers in parentheses represent the mean plausibility rating for that answer choice. The answer choice with the highest plausibility rating is underlined.

The advantages of MCQ evaluation are clear: with a single correct choice per question, system scores are easy to compute and understand. However, by their nature, commonsense reasoning questions typically do not have a single objectively correct answer; rather they admit many possible answers with varying degrees of plausibility as shown in Figure 1. Under these conditions, what does it mean for a commonsense MCQ answer choice to be the "correct" answer?

We posit that the "correct" MCQ answer in this setting should be the one that human annotators agree is *most plausible* among options. In principle, the plausibility of an individual MCQ answer choice should depend only on the MCQ context (if applicable), question, and the answer choice itself, but need *not* depend on the other answer choices.[2] Under this assumption, then, a valid procedure to determine the correct MCQ answer would be to rate the plausibility of each choice individually and select the highest-scoring option.

In this paper, we analyze two important commonsense MCQ benchmarks, Social IQa (SIQA; Sap et al., 2019) and CommonsenseQA (CSQA; Tal-

[1]Our data is available at https://github.com/shramay-palta/commonsense-mcq-plausibility

[2]An obvious exception is if an answer choice directly refers to other options, e.g. "None of the above."



*Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3451–3473
November 12-16, 2024 ©2024 Association for Computational Linguistics

| Dataset | #MCQ samples (#Answers) | #Full Anno. (#Tie Break) | #Plaus. Ratings |
|---|---|---|---|
| SIQA | 125(375) | 765(140) | 1875 |
| CSQA | 125(625) | 765(140) | 3125 |
| Total | **250**(1000) | **1530**(280) | **5000** |

Table 1: Number of annotations performed on Social IQa (SIQA) and CommonsenseQA (CSQA) samples for the tasks of Individual Plausibility Rating (§ 2.1) and Full Question Annotation (§ 2.2); totals are bolded.

mor et al., 2019), through the lens of this individual plausibility rating procedure. On 250 MCQ items sampled from both datasets, we collect 5 Likert-scale plausibility ratings of individual answers in isolation (§ 2.1), and 5-10 best answer choice judgments given the full set of answers (§ 2.2). With this data, we are able to make the following observations and conclusions:

1. While gold answers for MCQs receive the highest average plausibility rating in a large majority of cases, we observe that, surprisingly, the gold and most-plausible answers do not align in over 20% of sampled MCQs for both datasets.
2. Through a qualitative analysis of these instances where gold and most-plausible answers do not align, we find a high prevalence of issues such as question ambiguity and answer choices that do not fit the question, among others.
3. MCQs in which the *difference* in mean plausibility scores between the most plausible and second-most plausible answer choices is small are more likely to exhibit low agreement on best answer choice judgments (§ 3).
4. Experiments with LLMs reveal low accuracy and high variation in performance (§ 4) on these instances, indicating our approach can help to identify more reliable benchmark items for commonsense evaluation.

## 2 Human Data Collection

We select CSQA (Talmor et al., 2019) and SIQA (Sap et al., 2019) for our study as they are popular MCQ benchmarks for general commonsense and social commonsense reasoning, respectively.

**Social IQa**: MCQ items consist of a short context describing a social situation, a question about a person in the situation, and three answer choices (see Fig. 1.) We randomly sample 125 questions from the `validation` split. These MCQ items were originally assigned a gold answer choice based on a majority vote of five annotators.

| Statistic | SIQA | CSQA |
|---|---|---|
| Original Gold Answer | 3.86(0.73) | 4.23(0.71) |
| Maximum Rating | 3.98(0.67) | 4.33(0.63) |
| Second-Best Rating | 2.88(0.74) | 3.23(0.99) |
| Minimum Rating | 2.12(0.67) | 1.43(0.47) |
| Maximum - Second-Best | 1.10(0.77) | 1.10(0.83) |
| Maximum - Minimum | 1.86(0.83) | 2.90(0.67) |

Table 2: Mean Likert-score for gold answers, most-plausible answers, second-most plausible answers, least-plausible answers, and average differences. Numbers in parentheses represent the standard deviation.

**CommonsenseQA**: MCQ items consist of a question generated by humans using CONCEPTNET (Speer et al., 2017) relations and five possible answer choices. We sample another 125 validation questions, which have gold labels based on approval by a second annotator after construction.

For each of the 250 sampled MCQ items from these two datasets, we collect two types of human judgments: **individual plausibility ratings** (§ 2.1) and **full question annotations** (§ 2.2). Annotators are recruited through Prolific and paid $15/hour; see Appendix A.4 for details, including annotation interfaces (Figure 5 and Figure 6). Annotation counts for the two tasks are presented in Table 1.

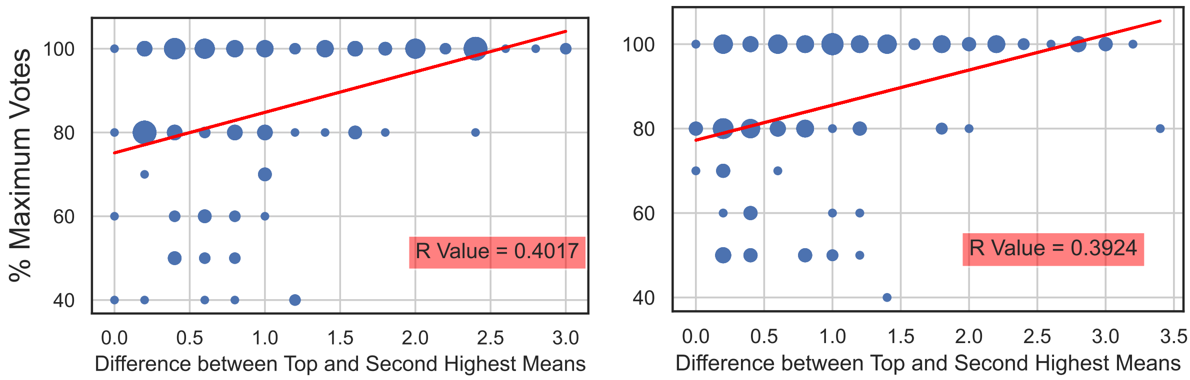
### 2.1 Individual Plausibility Ratings

To obtain the plausibility ratings for each option for a given question, we break down each question $q$ with choices $c_1, c_2, ...c_n$ into pairs $(q, c_i)$, where $n = 3$ for SIQA and 5 for CSQA.

Each $(q, c_i)$ tuple is presented to annotators where they are instructed to "rate the plausibility of the answer choice for the given question on a 5-point Likert scale". We use the plausibility Likert scale introduced by Zhang et al. (2017) for ordinal common-sense inference, defined as ***1-Impossible****, **2-Technically Possible**, **3-Plausible**, **4-Likely** and **5-Very Likely**.*

We obtain 5 annotations for each $(q, c_i)$ tuple. To ensure independence, each annotator judges at most one $(q, c_i)$ tuple for a given question $q$. Krippendorff's $\alpha$ on SIQA and CSQA is 0.46 and 0.64, respectively.[3]

[3]We hypothesise that the low Krippendorff's alpha value for SIQA in § 2.1 is due to the higher difficulty in judging the plausibility of social situations from ATOMIC (SIQA) as compared to commonsense knowledge in ConceptNet (CSQA).




(a) (b)

Figure 2: Difference in the plausibility scores between the top 2 most plausible choices (§ 2.1) vs. percentage of votes (§ 2.2) received by the top choice (on SIQA (left) and CSQA (right)). The size of the point represents the number of data points at an instance.

For each $(q, c_i)$ tuple, we compute the mean plausibility rating. Mean plausibility statistics are reported in Table 2.

### 2.2 Full Question Annotation

In this setting, annotators are provided the full MCQ item with all the answer choices and asked to select the (single) best option, similar to the validation procedures used to obtain original gold labels. However, to measure human agreement, we re-collect these annotations ourselves in larger numbers. Each MCQ item first receives five annotations; if no answer choice receives a majority vote from the annotators by a margin of two or more, then five more annotations are collected for the item. Krippendorff's $\alpha$ on SIQA and CSQA is $0.66$ and $0.71$, respectively. In over $87\%$ of cases on both datasets, the majority vote from our annotators matches the original gold label in the datasets.

## 3 Plausibly Problematic MCQs

With these collected judgments, we consider three ways to define a "correct" answer choice for each MCQ item: (1) the original gold answer choices from SIQA or CSQA ($y_{dataset}$), (2) the majority-vote answer choice from full question annotation ($y_{full}$), and (3) the answer choice with the maximum mean plausibility rating ($y_{plausibility}$). We hypothesize that $y_{plausibility}$ should be predictive of $y_{dataset}$ and $y_{full}$ across MCQs, and that when they diverge it may be indicative of one or more problems with the underlying MCQ.

To corroborate this idea, first we show in Figure 2a and Figure 2b that a small difference in plausibility scores between the highest- and second-highest scoring answers in the individual plausibility setting is correlated with lower agreement on the full question annotations, for both datasets.[4] This is consistent with the idea that disagreements on full MCQ annotations may arise when there is not a clear most-plausible answer.

Next we compare $y_{plausibility}$ to $y_{dataset}$. For both SIQA and CSQA, $y_{plausibility}$ diverges from $y_{dataset}$ in $22.4\%$ of MCQs. We define these MCQs as "plausibly problematic" questions given that the answer choice selected as $y_{plausibility}$ did not match $y_{dataset}$.

### 3.1 Qualitative Analysis

We conduct a manual inspection to identify the key issues with these "plausibly problematic" questions (identified using the plausibility judgements from § 2.1) and examine all such questions from SIQA and CSQA. We categorize the potential issues as: **1) Semantic Mismatch or Constraints:** A semantic discrepancy exists either between the question and at least one answer choice, or the question implies specific semantic limitations that at least one answer choice fails to meet; **2) Question is not coherent:** The question is not properly structured, leading to confusion and lack of clarity or is a poor fit for the context;[5] **3) Ambiguous:** The question requires one or more implicit assumptions to pick an answer (see Figure 1); **4) No good answer choices:** There are no answer choices that

[4]The p-value for SIQA and CSQA was evaluated to be $3.42E^{-06}$ and $6E^{-06}$ respectively.

[5]For SIQA, we concatenate the context and question as the question, as CSQA has no 'context' field.

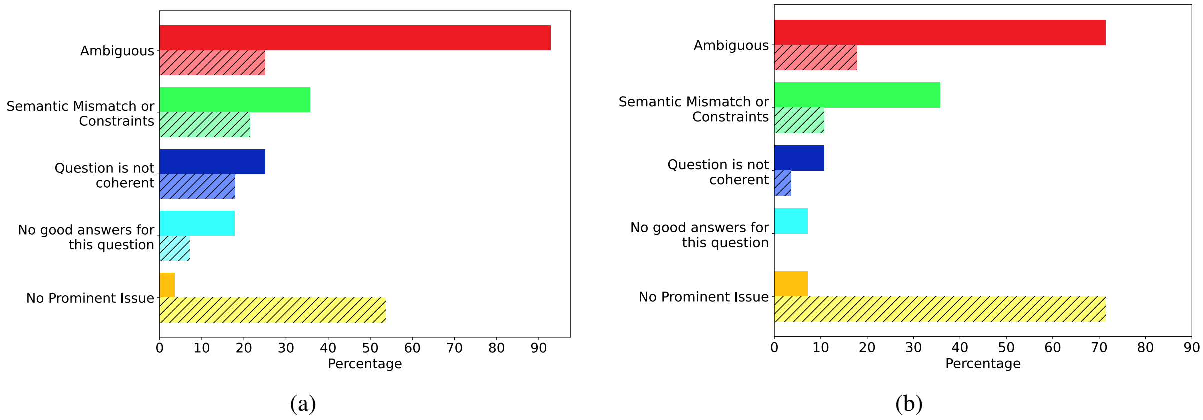



Figure 3: Frequency of issues types on the “plausibly problematic” (solid) and non-problematic (hatched) questions from SIQA (left) and CSQA (right) (28 MCQs each). It is important to note that these labels are not mutually exclusive and a question can be “plausibly problematic” due to multiple reasons and hence tagged with more than one label.

are a good fit for the question; and **5) No Prominent Issue:** There is no prominent issue with the question. Examples of questions with each of these labels are presented in Table 9 in Appendix A.5.

As seen in Figure 3a and Figure 3b, **Ambiguous** and **Semantic Mismatch or Constraints** are the most common issues with the “plausibly problematic” questions. The prevalence of these labels indicates that there are questions in both of these datasets which have multiple possible valid interpretations. We recommend future works to build upon our findings and urge dataset creators to ensure that the questions in their datasets do not have multiple different but valid interpretations, and that all answer choices should be geared towards one interpretation. We also encourage dataset creators to include of “not applicable” or “question does not make sense” option (Dowty, 1991), especially when creating datasets involving automatic assignments of questions.

We also observe very few cases where a question is tagged with the **No Prominent Issue** label, which could be attributed to noise from the human annotations (§ 2.1). A similar analysis on an equal number of questions sampled randomly from the set of “Non-Problematic” Questions is presented in Figure 3a and Figure 3b. We find that a vast majority of the non-problematic questions would receive the **No Prominent Issue** label, suggesting that the questions were clear and had an answer choice which was clearly suited better than the others. This indicates that our approach is also able to identify non-problematic questions accurately.

## 4 Implications for LLM Evaluation

We prompt LLMs with the same task posed to humans in § 2.1 and § 2.2. We study multiple state-of-the-art LLMs: GPT-4 (`gpt-4-0125-preview`) (Achiam et al., 2023) with the OpenAI API, LLaMA-2 (7B, 13B and 70B) (Touvron et al., 2023), Mistral (7B and 7x8B) (Jiang et al., 2024) and Yi (6B, 9B and 34B) (AI et al., 2024). We prompt each LLM with the same 10 in-context examples for the Plausibility and Full settings.[6]

| **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 |

Table 3: Percentages of cases where the agent response in the full question setting matches the original dataset gold label on the set of “plausibly problematic” and non-problematic questions (identified using plausibility judgements from § 2.1) from SIQA and CSQA.

We compare human and LLM performance on the set of “plausibly problematic” and non-problematic questions (identified using the plausibility ratings (§ 2.1)) and present the accuracy

[6]We present the questions used for in-context examples in Appendix A.

(against $y_{dataset}$) in Table 3. We observe that (1) accuracy on the "plausibly problematic" subset is lower, and (2) the performance drop in the problematic set is larger for LLMS than for humans. The overall lower performance on the "plausibly problematic" subset also suggests that these questions are not merely hard to answer for the models, but have certain underlying issues associated with them, which we discussed in § 3.1.

## 5 Related Works

**Dataset Quality Analysis:** Many works find biases in datasets, including dataset artifacts (Poliak et al., 2018; Gururangan et al., 2018; Balepur et al., 2024b; Balepur and Rudinger, 2024) and annotator noise (Sheng et al., 2008; Snow et al., 2008; Nowak and Rüger, 2010). Given these findings, recent work has proposed not to treat every data entry as equally difficult when assessing LMs (Rodriguez et al., 2021), using human psychology techniques such as Item Response Theory (Lalor et al., 2016; Vania et al., 2021; Rodriguez et al., 2022) or model-based hardness metrics (Perez et al., 2021). Swayamdipta et al. (2020) use this method to disentangle difficult and ambiguous/noisy data entries. Similarly, we show how plausibility ratings can uncover problematic data in MCQ datasets.

**Plausibility in Commonsense:** Ranking, comparing, and scoring the plausibility of events and outcomes expressed in language is a longstanding concept in commonsense reasoning research(Roemmele et al., 2011; Wang et al., 2018; Li et al., 2019; Liu et al., 2023). Because commonsense knowledge is often subjective (Whiting and Watts, 2024) or graded (Zhang et al., 2017; Chen et al., 2020), and varies with cultural context (Palta and Rudinger, 2023; Hershcovich et al., 2022; Bhatia and Shwartz, 2023), this can pose challenges for evaluation. Most relevant to this work, Acquaye et al. (2024) use Likert-scale human plausibility judgments of answer choices to construct cultural commonsense MCQ test items. Other approaches to evaluation include verbalized rationales (Jung et al., 2022; Balepur et al., 2024a). Specifically, prior works have studied defeasible (Rudinger et al., 2020; Rao et al., 2023) and abductive reasoning (Bhagavatula et al., 2020) in natural language, where models rationalize when scenarios may be more plausible or valid.

## 6 Conclusion

In this work, we show that plausibility judgments are a useful tool for identifying *potentially* problematic commonsense MCQ items. With individual plausibility ratings, we are able to identify questions where the gold answer does not match the answer with the highest plausibility. Through manual analysis we identify several types of issues that are more prevalent among the identified subset. We show that LLMS and humans perform poorly on these questions, with a high degree of variance, suggesting they add noise to benchmark evaluations. Future work may investigate methods of incorporating plausibility judgments into the creation stage of benchmark development, as well as the application of these ideas to evaluating other types of benchmarks involving graded judgments beyond commonsense reasoning.

## 7 Limitations

Uncertainty can arise due to a variety of reasons such as multi-cultural and multi-ethnic aspects of commonsense reasoning. In this work while we introduce a new method to identify questions with multiple plausible answers, we are limited to a US-centric angle of uncertainty owing to the fact that our annotators are based in the US.

Additionally, our annotation framework is expensive and thus difficult to run on an entire dataset. However, since we are the first to explore plausibility of answer choices in commonsense reasoning situations, we hope that this work motivates other researchers to study plausibility more extensively.

The identification and annotation of uncertainty can be subjective, leading to inconsistencies or disagreements among annotators. While we employed rigorous annotation protocols and made sure each question was annotated by at least 5 annotators, there may still be instances where ambiguity interpretation varies.

## 8 Acknowledgements

We would like to thank the anonymous reviewers for their valuable feedback on this paper. We would also like to thank Hal Daumé III, Faeze Brahman, Elijah Rippeth, Alexander Hoyle, Yuelin Liu, Sander Schulhoff, Abhilasha Sancheti, Haozhe An, Neha Srikanth, Christabel Acquaye, Yu Hou and other members of the CLIP lab for for their helpful comments and suggestions. Nishant Balepur's funding and this material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE 2236417. Rachel Rudinger is supported by NSF CAREER Award No. 2339746. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

## References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Christabel Acquaye, Haozhe An, and Rachel Rudinger. 2024. Susu box or piggy bank: Assessing cultural commonsense knowledge between ghana and the us. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami. Association for Computational Linguistics.

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.ai.

Nishant Balepur, Shramay Palta, and Rachel Rudinger. 2024a. It's not easy being wrong: Large language models struggle with process of elimination reasoning. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10143–10166, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Nishant Balepur, Abhilasha Ravichander, and Rachel Rudinger. 2024b. Artifacts or abduction: How do LLMs answer multiple-choice questions without the question? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10308–10330, Bangkok, Thailand. Association for Computational Linguistics.

Nishant Balepur and Rachel Rudinger. 2024. Is your large language model knowledgeable or a choices-only cheater? In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 15–26, Bangkok, Thailand. Association for Computational Linguistics.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen Tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *8th International Conference on Learning Representations, ICLR 2020*.

Mehar Bhatia and Vered Shwartz. 2023. GD-COMET: A geo-diverse commonsense inference model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7993–8001, Singapore. Association for Computational Linguistics.

Yonatan Bisk, Rowan Zellers, Ronan Le bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7432–7439.

Tongfei Chen, Zhengping Jiang, Adam Poliak, Keisuke Sakaguchi, and Benjamin Van Durme. 2020. Uncertain natural language inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8772–8779, Online. Association for Computational Linguistics.

David Dowty. 1991. Thematic proto-roles and argument selection. *language*, 67(3):547–619.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. 2022. Maieutic prompting: Logically consistent reasoning with recursive explanations. *arXiv preprint arXiv:2205.11822*.

John P. Lalor, Hao Wu, and Hong Yu. 2016. Building an evaluation scale using item response theory. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 648–657, Austin, Texas. Association for Computational Linguistics.

Zhongyang Li, Tongfei Chen, and Benjamin Van Durme. 2019. Learning to rank for plausible plausibility. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4818–4823, Florence, Italy. Association for Computational Linguistics.

Jiacheng Liu, Wenya Wang, Dianzhuo Wang, Noah Smith, Yejin Choi, and Hannaneh Hajishirzi. 2023. Vera: A general-purpose plausibility estimation model for commonsense statements. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1264–1287, Singapore. Association for Computational Linguistics.

Stefanie Nowak and Stefan Rüger. 2010. How reliable are annotations via crowdsourcing? a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on Multimedia information retrieval - MIR '10*, page 557. This is the author's version of the work. It is posted here by permission of ACM for your personal use. Not for redistribution.

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*, pages 9952–9962, Toronto, Canada. Association for Computational Linguistics.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. Rissanen data analysis: Examining dataset characteristics via description length. In *International Conference on Machine Learning*, pages 8500–8513. PMLR.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Kavel Rao, Liwei Jiang, Valentina Pyatkin, Yuling Gu, Niket Tandon, Nouha Dziri, Faeze Brahman, and Yejin Choi. 2023. What makes it ok to set a fire? iterative self-distillation of contexts and rationales for disambiguating defeasible social and moral situations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12140–12159, Singapore. Association for Computational Linguistics.

Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. 2021. Evaluation examples are not equally informative: How should that change NLP leaderboards? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4486–4503, Online. Association for Computational Linguistics.

Pedro Rodriguez, Phu Mon Htut, John Lalor, and João Sedoc. 2022. Clustering examples in multi-dataset benchmarks with item response theory. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 100–112, Dublin, Ireland. Association for Computational Linguistics.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith, and Yejin Choi. 2020. Thinking like a skeptic: Defeasible inference in natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4661–4675, Online. Association for Computational Linguistics.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, page 614–622, New York, NY, USA. Association for Computing Machinery.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Clara Vania, Phu Mon Htut, William Huang, Dhara Mungra, Richard Yuanzhe Pang, Jason Phang, Haokun Liu, Kyunghyun Cho, and Samuel R. Bowman. 2021. Comparing test sets with item response theory. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1141–1158, Online. Association for Computational Linguistics.

Su Wang, Greg Durrett, and Katrin Erk. 2018. Modeling semantic plausibility by injecting world knowledge. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 303–308.

Mark E Whiting and Duncan J Watts. 2024. A framework for quantifying individual and collective common sense. *Proceedings of the National Academy of Sciences*, 121(4):e2309535121.

Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017. Ordinal common-sense inference. *Transactions of the Association for Computational Linguistics*, 5:379–395.

# A Appendix

## A.1 License for Artifacts

All datasets used in this work are publicly available and free to use on HuggingFace.

## A.2 Details on Computational Experiments

LLaMA-2 70B, LLaMA-2-13B, Yi-34B, Yi-9B, and Mixtral 7x8B were all run on eight NVIDIA:RTXA5000 GPUs and were allocated a total of eight GPU hours to run all experiments. All other open-source LLMs were run on one NVIDIA:RTXA6000 GPU and were allocated a total of two GPU hours to run all experiments. GPT-4 was run on CPU and was allocated one hour to run all experiments. Each LLM decodes with a minimum token generation length of 5, a maximum token generation length of 200, greedy decoding (or 0 temperature in the case of GPT-4), and a stopping criteria when the LLM begins to generate the next few-shot exemplar. We did not perform a hyperparameter search. All results are obtained from a single run.

| **Agent** | **SIQA** | **CSQA** |
|---|---|---|
| LLaMA-2 7B | 38.3 | 22.7 |
| LLaMA-2 13B | 53.3 | 56.4 |
| LLaMA-2 70B | 60.1 | 58.9 |
| Mistral 7B | 44.5 | 56.6 |
| Mixtral 7x8B | 68.6 | 58.6 |
| Yi 6B | 44.8 | 31.6 |
| Yi 9B | 66.2 | 55.9 |
| Yi 34B | 73.9 | 64.5 |
| GPT-4 | 73.0 | 69.4 |
| Average LLM | 58.1 | 52.7 |
| Human | 77.9 | 77.2 |

Table 4: Percentage of cases where the most plausible answer from § 2.1 matches the response to the full question from § 2.2.

## A.3 Additional Experiments and Results

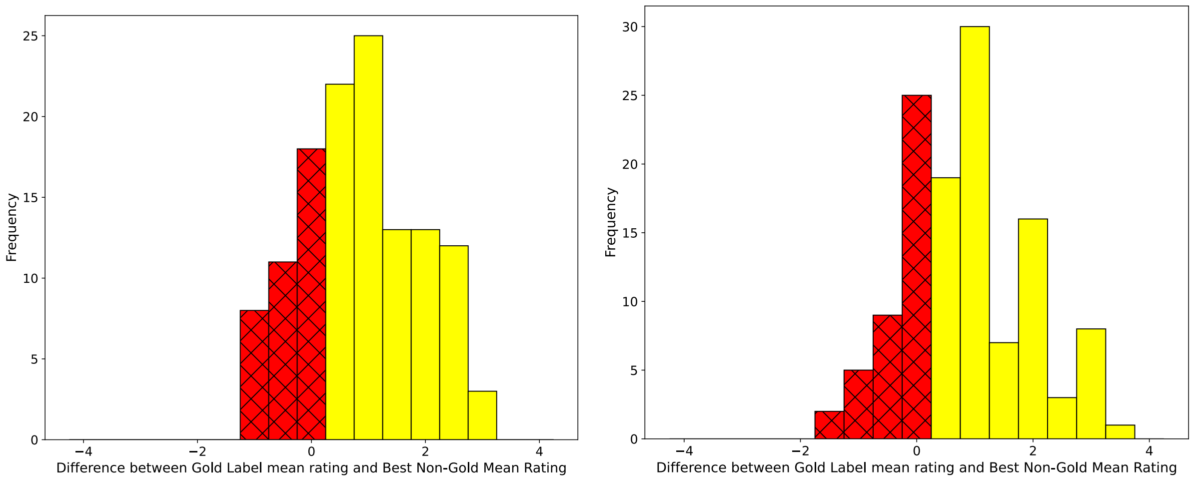
### A.3.1 Full Question Setting

In this setting, for humans, we look at the vote distribution for each question and use that to determine whether $y_{full} = y_{dataset}$. We flag the questions as "problematic" in the *Full Question Setting* if $y_{full} \neq y_{dataset}$ or the difference between the highest and second highest votes (for humans) is less than 2.

We observe that in the *Full Question setting*, humans exhibit overall better performance than LLMs (highlighted in Table 7), suggesting that even when presented with all the answer choices, 'problematic' questions pose a challenge to effective LLM evaluation of commonsense reasoning capabilities.

Table 4 demonstrates the cases where $y_{plausibility} = y_{full}$. Table 5 shows the Pearson's Correlation Coefficient for Human and LLM individual plausibility ratings.

Table 6 and Table 8 demonstrate that LLMs show higher agreement with the human responses in cases where the questions are not identified as problematic. This finding is consistent in both the *Individual Plausibility Setting* (§ 2.1) and the *Full Question Setting* (§ 2.2).




(a) Social IQa (b) CommonsenseQA

Figure 4: Histograms showing the difference between the mean gold label rating and best non-gold label rating. Portions of the graph in red with texture show cases where the best non-gold option had a higher mean plausibility rating than the mean gold label rating.

| Model | SIQA | CSQA |
|---|---|---|
| LLaMA-2 7B | 0.262 | 0.178 |
| LLaMA-2 13B | 0.417 | 0.654 |
| LLaMA-2 70B | 0.656 | 0.760 |
| Mistral 7B | 0.385 | 0.675 |
| Mixtral 7x8B | 0.573 | 0.709 |
| Yi 6B | 0.330 | 0.399 |
| Yi 9B | 0.652 | 0.700 |
| Yi 34B | 0.648 | 0.716 |
| GPT-4 | 0.708 | 0.775 |

Table 5: Pearson's Correlation coefficients between LLM plausibility ratings and human plausibility ratings.

## A.4 Annotation Process Details

We used Prolific to collect the human annotations. The annotators for our task were selected on the basis of the following criteria:

1. Must be located in the United States.
2. Primary language must be English.
3. Must not have any literacy difficulties.
4. Must have attained a minimum of an undergraduate level degree.
5. Must have an approval rate between $95 - 100\%$ on Prolific.
6. We use a $50 - 50$ split of male and female[7] annotators to minimize the risk of any gender-specific biases creeping in.

The total cost for our entire annotation protocols, for both Individual Plausibility Ratings, and the Full Question Setting came out to be $1052. We also received an exempt status from the IRB at our institution for this research.

## A.5 Examples of Questions with Labels

We include an example question for each label used in our error analysis as described in § 3.1 and present them in Table 9.

[7] Gender as indicated on Prolific.

| Model | SIQA Problematic | SIQA Non-Problematic | SIQA Overall | CSQA Problematic | CSQA Non-Problematic | CSQA Overall |
|---|---|---|---|---|---|---|
| LLaMA-2 7B | 61.5 | 67.4 | 66.1 | 70.4 | 68.2 | 68.7 |
| LLaMA-2 13B | 51.9 | 72.5 | 67.8 | 63.0 | 75.0 | 72.2 |
| LLaMA-2 70B | 50.0 | 87.6 | 79.1 | 81.5 | 83.5 | 83.0 |
| Mistral 7B | 48.1 | 78.7 | 71.7 | 74.1 | 77.8 | 77.0 |
| Mixtral 7x8B | 51.9 | 84.3 | 77.0 | 77.8 | 85.8 | 83.9 |
| Yi 6B | 53.8 | 82.6 | 76.1 | 66.7 | 79.0 | 76.1 |
| Yi 9B | 59.6 | 89.9 | 83.0 | 77.8 | 83.5 | 82.2 |
| Yi 34B | 55.8 | 94.4 | 85.7 | 77.8 | 91.5 | 88.3 |
| GPT-4 | 59.6 | 88.8 | 82.2 | 74.1 | 88.1 | 84.8 |

Table 6: Instances where the LLM response matches the response given by humans, based on maximum vote on the set of "plausibly problematic" and non-problematic questions (identified from the *Individual Plausibility Rating Setting*) in the Full Question Setting.

| Agent | SIQA Problematic | SIQA Non-Problematic | SIQA Overall | CSQA Problematic | CSQA Non-Problematic | CSQA Overall |
|---|---|---|---|---|---|---|
| LLaMA-2 7B | 45.5 | 66.3 | 62.9 | 33.3 | 66.1 | 62.1 |
| LLaMA-2 13B | 45.5 | 70.2 | 66.1 | 33.3 | 74.3 | 69.3 |
| LLaMA-2 70B | 45.5 | 84.6 | 78.1 | 33.3 | 83.5 | 77.4 |
| Mistral 7B | 45.5 | 74.0 | 69.3 | 33.3 | 74.3 | 69.3 |
| Mixtral 7x8B | 45.5 | 82.7 | 76.6 | 50.0 | 84.4 | 80.2 |
| Yi 6B | 45.5 | 79.8 | 74.1 | 66.7 | 79.8 | 78.2 |
| Yi 9B | 72.7 | 88.5 | 85.9 | 50.0 | 84.4 | 80.2 |
| Yi 34B | 81.8 | 87.5 | 86.6 | 50.0 | 88.1 | 83.5 |
| GPT-4 | 54.5 | 84.6 | 79.6 | 50.0 | 86.2 | 81.8 |
| Average LLM | 53.6 | 79.8 | 75.5 | 44.4 | 80.1 | 75.8 |
| Human | 22.7 | 96.2 | 84.1 | 8.3 | 91.7 | 81.5 |

Table 7: Percentages of cases where agent response to the full question matches the original dataset gold label on the set of problematic and non-problematic questions (identified from the *Full Question setting*) from SIQA and CSQA.

| Model | SIQA Problematic | SIQA Non-Problematic | SIQA Overall | CSQA Problematic | CSQA Non-Problematic | CSQA Overall |
|---|---|---|---|---|---|---|
| LLaMA-2 7B | 27.3 | 70.2 | 63.1 | 33.3 | 70.6 | 66.1 |
| LLaMA-2 13B | 27.3 | 72.1 | 64.7 | 50.0 | 73.4 | 70.6 |
| LLaMA-2 70B | 27.3 | 84.6 | 75.1 | 58.3 | 84.4 | 81.2 |
| Mistral 7B | 31.8 | 76.0 | 68.7 | 58.3 | 78.0 | 75.6 |
| Mixtral 7x8B | 22.7 | 82.7 | 72.8 | 41.7 | 86.2 | 80.8 |
| Yi 6B | 22.7 | 81.7 | 72.0 | 25.0 | 78.9 | 72.3 |
| Yi 9B | 31.8 | 88.5 | 79.1 | 41.7 | 84.4 | 79.2 |
| Yi 34B | 31.8 | 91.3 | 81.5 | 41.7 | 90.8 | 84.8 |
| GPT-4 | 40.9 | 86.5 | 79.0 | 41.7 | 87.2 | 81.6 |

Table 8: Instances where the LLM response matches the response given by humans, based on maximum vote on the set of problematic and non-problematic questions (identified from the *Full Question Setting*) in the Full Question Setting.

| Label | Social IQA | CommonsenseQA |
|---|---|---|
| Ambiguous | *Context*: After seeing what a mess Aubrey was, Robin changed her into clean clothes.<br>*Question*: How would you describe Robin?<br>*Choices*:<br>(A) **a kind caretaker**<br>(B) like a person who puts in thought<br>(C) <u>a reliable friend</u><br>*Explanation:* One needs to assume the relationship between Aubrey and Robin to be able to pick a response. | *Question*: When you get together with friends to watch film, you might do plenty of this?<br>*Choices*:<br>(A) <u>see what happens</u><br>(B) enjoy stories<br>(C) pass time<br>(D) **have fun**<br>(E) interesting<br>*Explanation:* Answers B, C and D are all acceptable responses. Answer A does not specify what one is actually "seeing". |
| Semantic Mismatch or Constraint | *Context*: Jesse just got a haircut and Riley was observing him with her eyes.<br>*Question*: What will happen to Jesse?<br>*Choices*:<br>(A) **Give a compliment to Jesse about his hair**<br>(B) <u>go for a haircut</u><br>(C) see Jesse's haircut<br>*Explanation:* None of the answer choices describe an event that can "happen" to Jesse. | *Question*: What regions of a town would you have found a dime store?<br>*Choices*:<br>(A) <u>commercial building</u><br>(B) old movie<br>(C) **small neighborhood**<br>(D) past<br>(E) mall<br>*Explanation:* Answers B and D are not "regions of a town". |
| Question is not coherent | *Context*: Remy answered the silly question they were asked happily.<br>*Question*: Why did Remy do this?<br>*Choices*:<br>(A) know the answer<br>(B) <u>think about fun</u><br>(C) **have fun**<br>*Explanation*: The question does not ask about anything mentioned in the context. None of the answer choices are a suitable response to the question. | *Question*: The flower grew tall to compete for sunlight, what did its neighbor do?<br>*Choices*:<br>(A) <u>blossom</u><br>(B) park<br>(C) open<br>(D) **cast shadow**<br>(E) vase<br>*Explanation*: The question does not mention who "neighbor" refers to. |
| No good answer choices | *Context*: Skylar wasn't certain that they had turned off the stove, so they went back to check.<br>*Question*: What does Skylar need to do before this?<br>*Choices*:<br>(A) **anxious**<br>(B) <u>needed to have turned on the toaster</u><br>(C) good<br>*Explanation*: None of the answer choices are a suitable response to the question. | *Question*: What would a person need to do if his or her captain dies at sea?<br>*Choices*:<br>(A) cross street<br>(B) have a party<br>(C) <u>experience life</u><br>(D) cross road<br>(E) **man crew**<br>*Explanation*: None of the answer choices are a suitable response to the question. |
| No prominent issue | *Context*: Robin had a hard time understanding the concept, so she let Carson explain it more thoroughly.<br>*Question*: How would Carson feel as a result?<br>*Choices*:<br>(A) **frustrated that Robin didn't understand**<br>(B) ready to play<br>(C) <u>ready to work</u> | *Question*: What are people likely to do when an unexpected decent outcome occurs?<br>*Choices*:<br>(A) kill each other<br>(B) **thank god**<br>(C) experience pain<br>(D) hatred<br>(E) <u>talk to each other</u> |

Table 9: Examples of "plausibly problematic" questions from SIQA and CSQA with labels. Text in blue (also underlined) indicates $y_{plausibility}$ and text in red (also bolded) indicates $y_{dataset}$. It is important to note that these labels are not mutually exclusive and a question can be "plausibly problematic" due to multiple reasons. Some of the above questions were tagged with more than one label, but we present unique questions for each label above.

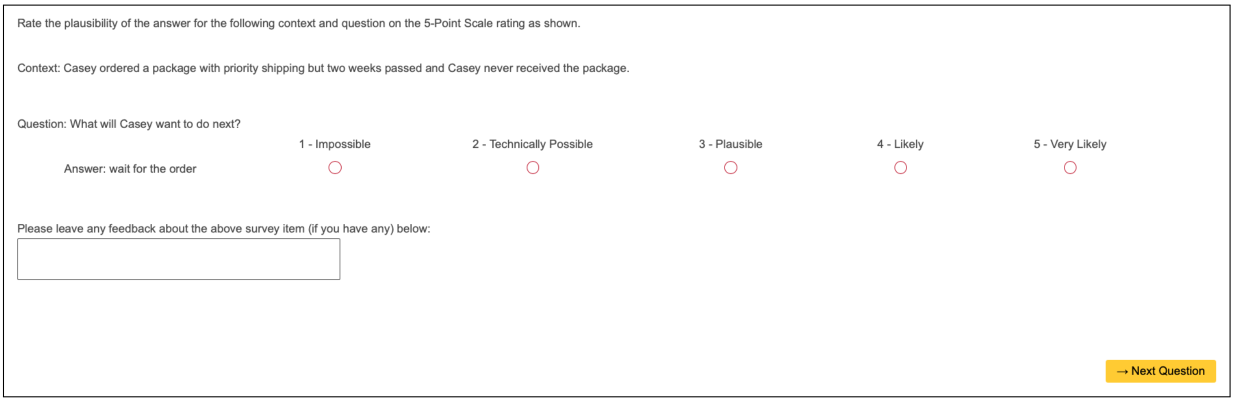



Figure 5: An example of the interface that annotators used while giving plausibility ratings to answer choices as described in § 2.1.

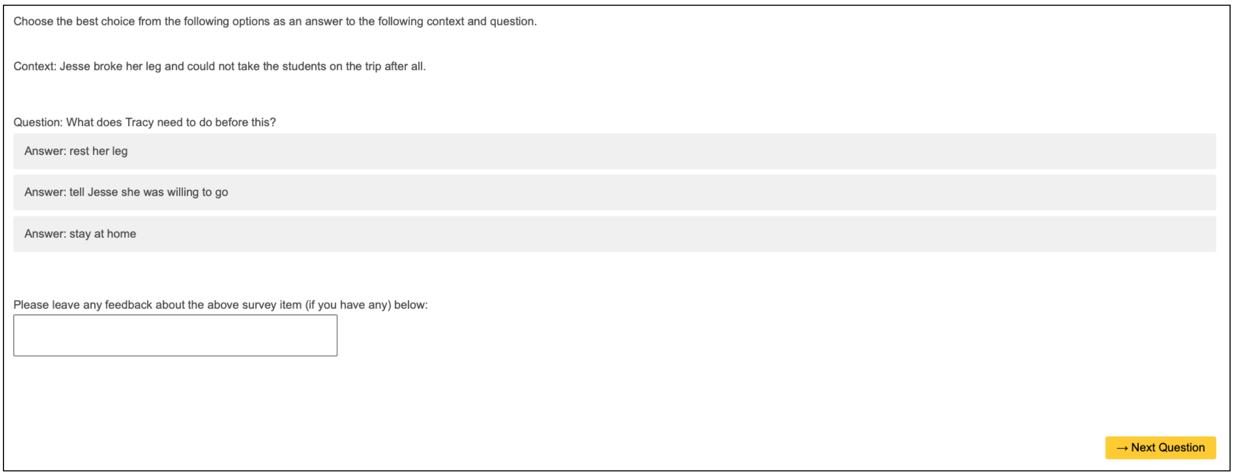



Figure 6: An example of the interface that annotators used while choosing the best answer choice for a question as described in § 2.2.

Context: Quinn got a new job in a different town so they moved to a new house.
Question: What will Quinn want to do next?
Choice: unpack his things
Plausibility Ratings:
(1) Impossible
(2) Technically Possible
(3) Plausible
(4) Likely
(5) Very Likely
Rating: (5)

Context: Remy was telling a story and told it like Casey does.
Question: Why did Remy do this?
Choice: tell Casey's story
Plausibility Ratings:
(1) Impossible
(2) Technically Possible
(3) Plausible
(4) Likely
(5) Very Likely
Rating: (4)

Context: Jesse showed Sydney's friends how to bake a cake but it was left in the oven too long and burned.
Question: What will Jesse want to do next?
Choice: Give the cake to friends
Plausibility Ratings:
(1) Impossible
(2) Technically Possible
(3) Plausible
(4) Likely
(5) Very Likely
Rating: (2)

Context: Skylar wasn't certain that they had turned off the stove, so they went back to check.
Question: What does Skylar need to do before this?
Choice: good
Plausibility Ratings:
(1) Impossible
(2) Technically Possible
(3) Plausible
(4) Likely
(5) Very Likely
Rating: (1)

Figure 7: In-context learning examples from Social IQa for the isolated setting. (Part 1)

Context: After graduating high school and being accepted to college, Casey decided to study philosophy, because Casey valued ethics highly.
Question: What does Casey need to do before this?
Choice: needed to register for the program
Plausibility Ratings:
(1) Impossible
(2) Technically Possible
(3) Plausible
(4) Likely
(5) Very Likely
Rating: (5)

Context: Sasha wetted the bed because they had a bad dream.
Question: How would Sasha feel afterwards?
Choice: has insecurities
Plausibility Ratings:
(1) Impossible
(2) Technically Possible
(3) Plausible
(4) Likely
(5) Very Likely
Rating: (3)

Context: Robin took the Math test at face value because it was based on logic.
Question: How would Robin feel afterwards?
Choice: lazy
Plausibility Ratings:
(1) Impossible
(2) Technically Possible
(3) Plausible
(4) Likely
(5) Very Likely
Rating: (2)

Context: Taylor looked the definition up online which was against the rules for the test.
Question: What will others want to do next?
Choice: had less preparation for the test
Plausibility Ratings:
(1) Impossible
(2) Technically Possible
(3) Plausible
(4) Likely
(5) Very Likely

Figure 8: In-context learning examples from Social IQa for the isolated setting. (Part 2)

Rating: (1)

Context: Ash saved the beehive from destruction because honey is good for you.
Question: How would Others feel as a result?
Choice: quite motivated
Plausibility Ratings:
(1) Impossible
(2) Technically Possible
(3) Plausible
(4) Likely
(5) Very Likely
Rating: (3)

Context: After years of bullying in high school, Sasha graduated at the top of her class and got a great job. Sasha got revenge on the people.
Question: How would you describe Sasha?
Choice: like they shouldn't of bullied her
Plausibility Ratings:
(1) Impossible
(2) Technically Possible
(3) Plausible
(4) Likely
(5) Very Likely
Rating: (4)

Figure 9: In-context learning examples from Social IQa for the isolated setting. (Part 3)

Context: Quinn got a new job in a different town so they moved to a new house.
Question: What will Quinn want to do next?
Choices:
(A) welcome
(B) unpack his things
(C) reload his things
Answer: (B)

Context: Remy was telling a story and told it like Casey does.
Question: Why did Remy do this?
Choices:
(A) learn casey
(B) tell Casey's story
(C) honor Casey
Answer: (C)

Context: Jesse showed Sydney's friends how to bake a cake but it was left in the oven too long and burned.
Question: What will Jesse want to do next?
Choices:
(A) teach everyone how to bake
(B) Give the cake to friends
(C) Bake a new cake
Answer: (C)

Context: Skylar wasn't certain that they had turned off the stove, so they went back to check.
Question: What does Skylar need to do before this?
Choices:
(A) anxious
(B) needed to have turned on the toaster
(C) good
Answer: (A)

Context: After graduating high school and being accepted to college, Casey decided to study philosophy, because Casey valued ethics highly.
Question: What does Casey need to do before this?
Choices:
(A) read every philosophy text he can find
(B) needed to register for the program
(C) become a philospher
Answer: (B)

Context: Sasha wetted the bed because they had a bad dream.
Question: How would Sasha feel afterwards?

Figure 10: In-context learning examples from Social IQa for the full setting. (Part 1)

Choices:
(A) has insecurities
(B) has a weak bladder
(C) the need to clean up right away
Answer: (C)

Context: Robin took the Math test at face value because it was based on logic.
Question: How would Robin feel afterwards?
Choices:
(A) smart
(B) unprepared
(C) lazy
Answer: (A)

Context: Taylor looked the definition up online which was against the rules for the test.
Question: What will others want to do next?
Choices:
(A) had less preparation for the test
(B) give Taylor a bad grade
(C) have made up his mind to cheat
Answer: (B)

Context: Ash saved the beehive from destruction because honey is good for you.
Question: How would Others feel as a result?
Choices:
(A) quite angry
(B) quite motivated
(C) a friend of the environment
Answer: (C)

Context: After years of bullying in high school, Sasha graduated at the top of her class and got a great job. Sasha got revenge on the people.
Question: How would you describe Sasha?
Choices:
(A) a competent person
(B) like they shouldn't of bullied her
(C) bad for bullying her
Answer: (A)

Figure 11: In-context learning examples from Social IQa for the full setting. (Part 2)

Question: What do people do when they don't understand something?
Choice: ask questions
Plausibility Ratings:
(1) Impossible
(2) Technically Possible
(3) Plausible
(4) Likely
(5) Very Likely
Rating: (5)

Question: What are people likely to do when an unexpected decent outcome occurs?
Choice: talk to each other
Plausibility Ratings:
(1) Impossible
(2) Technically Possible
(3) Plausible
(4) Likely
(5) Very Likely
Rating: (4)

Question: What do children require to grow up healthy?
Choice: fast food
Plausibility Ratings:
(1) Impossible
(2) Technically Possible
(3) Plausible
(4) Likely
(5) Very Likely
Rating: (2)

Question: The person wasn't bothered by the weather, she had remembered to bring her what?
Choice: own house
Plausibility Ratings:
(1) Impossible
(2) Technically Possible
(3) Plausible
(4) Likely
(5) Very Likely
Rating: (1)

Question: James knew that he shouldn't have been buying beer for minors. He didn't even get paid for it. Why was this bad?
Choice: broken law
Plausibility Ratings:

Figure 12: In-context learning examples from CommonsenseQA for the isolated setting. (Part 1)

(1) Impossible
(2) Technically Possible
(3) Plausible
(4) Likely
(5) Very Likely
Rating: (5)

Question: What would you do to a rock when climb up a cliff?
Choice: throw
Plausibility Ratings:
(1) Impossible
(2) Technically Possible
(3) Plausible
(4) Likely
(5) Very Likely
Rating: (3)

Question: What does everyone feel of monsters?
Choice: looking for love
Plausibility Ratings:
(1) Impossible
(2) Technically Possible
(3) Plausible
(4) Likely
(5) Very Likely
Rating: (2)

Question: Why does having a disability sometimes making academic tasks hard for a person?
Choice: acknowledgment
Plausibility Ratings:
(1) Impossible
(2) Technically Possible
(3) Plausible
(4) Likely
(5) Very Likely
Rating: (1)

Question: The teacher played on the upright piano, she was explaining the song to all the students in the what?
Choice: living room
Plausibility Ratings:
(1) Impossible
(2) Technically Possible
(3) Plausible

Figure 13: In-context learning examples from CommonsenseQA for the isolated setting. (Part 2)

(4) Likely
(5) Very Likely
Rating: (3)

Question: When getting in shape, this is something that does wonders?
Choice: period of recovery
Plausibility Ratings:
(1) Impossible
(2) Technically Possible
(3) Plausible
(4) Likely
(5) Very Likely
Rating: (4)

Figure 14: In-context learning examples from CommonsenseQA for the isolated setting. (Part 3)

Question: What do people do when they don't understand something?
Choices:
(A) believe in god
(B) experience joy
(C) ask questions
(D) talk to each other
(E) get sick
Answer: (C)

Question: What are people likely to do when an unexpected decent outcome occurs?
Choices:
(A) kill each other
(B) thank god
(C) experience pain
(D) hatred
(E) talk to each other
Answer: (B)

Question: What do children require to grow up healthy?
Choices:
(A) need care
(B) come home
(C) fast food
(D) watch television
(E) wash dishes
Answer: (A)

Question: The person wasn't bothered by the weather, she had remembered to bring her what?
Choices:
(A) read book
(B) own house
(C) apartment
(D) more rice
(E) warm coat
Answer: (E)

Question: James knew that he shouldn't have been buying beer for minors. He didn't even get paid for it. Why was this bad?
Choices:
(A) lose money
(B) fun
(C) have no money
(D) broken law
(E) relaxation

Figure 15: In-context learning examples from CommonsenseQA for the full setting. (Part 1)

Answer: (D)

Question: What would you do to a rock when climb up a cliff?
Choices:
(A) grab
(B) look down
(C) throw
(D) falling
(E) may fall
Answer: (A)

Question: What does everyone feel of monsters?
Choices:
(A) looking for love
(B) afraid of
(C) good at
(D) make pet
(E) different
Answer: (B)

Question: Why does having a disability sometimes making academic tasks hard for a person?
Choices:
(A) lots of space
(B) have choice
(C) mentally challenged
(D) hungry
(E) acknowledgment
Answer: (C)

Question: The teacher played on the upright piano, she was explaining the song to all the students in the what?
Choices:
(A) living room
(B) bathroom
(C) house
(D) music room
(E) music store
Answer: (D)

Question: When getting in shape, this is something that does wonders?
Choices:
(A) eat more
(B) starve
(C) give up

Figure 16: In-context learning examples from CommonsenseQA for the full setting. (Part 2)

(D) period of recovery
(E) jogging
Answer: (E)

Figure 17: In-context learning examples from CommonsenseQA for the full setting. (Part 3)