

Possible Stories: Evaluating Situated Commonsense Reasoning under Multiple Possible Scenarios

Mana Ashida*

Yahoo Japan Corporation
maashida@yahoo-corp.jp

Saku Sugawara

National Institute of Informatics
saku@nii.ac.jp

Abstract

The possible consequences for the same context may vary depending on the situation we refer to. However, current studies in natural language processing do not focus on situated commonsense reasoning under multiple possible scenarios. This study frames this task by asking multiple questions with the same set of possible endings as candidate answers, given a short story text. Our resulting dataset, Possible Stories, consists of more than 4.5K questions over 1.3K story texts in English. We discover that even current strong pretrained language models struggle to answer the questions consistently, highlighting that the highest accuracy in an unsupervised setting (60.2%) is far behind human accuracy (92.5%). Through a comparison with existing datasets, we observe that the questions in our dataset contain minimal annotation artifacts in the answer options. In addition, our dataset includes examples that require counterfactual reasoning, as well as those requiring readers' reactions and fictional information, suggesting that our dataset can serve as a challenging testbed for future studies on situated commonsense reasoning.

1 Introduction

Commonsense reasoning, inclusive of counterfactual, abductive, and monotonic reasoning, is a core element of language understanding. Researchers are interested in whether these abilities can be learned in systems, and several benchmarks have been proposed to investigate machine commonsense reasoning (Huang et al., 2019; Sap et al., 2019; Aggarwal et al., 2021; Saha et al., 2021). Recent pretrained models have shown competitive results (Khashabi et al., 2020; Lourie et al., 2021).

Commonsense reasoning has often been framed as a task to infer whether candidate answers are plausible, such as in the multiple-choice format (Talmor et al., 2019; Sakaguchi et al., 2020). The

*Work done while at Tokyo Metropolitan University.

Original Context

Cindy was planning to grow a lot of vegetables this year. She planted vegetable seedlings in her garden. Cindy knew there were hungry groundhogs in the area. She put up a short fence around her garden to protect it.

Original Ending

A: The groundhogs climbed over the fence and ate her seedlings.

Alternative Ending Collection

B: All of the ground animals were kept out, but something was still eating her vegetables.

C: She put spikes on the fence to avoid groundhogs and it worked.

D: No groundhogs climbed over the fence and Cindy had a good harvest in the fall.



Question Writing

Q1: Which one of the following is most likely to happen after this if there were other hungry animals? → Option B

Q2: What would be the most positive outcome for Cindy's crops? → Option D

Figure 1: Overview of Possible Stories and its creation process. We ask crowdworkers to produce alternative endings given a story text, and then write multiple-choice questions that have a single correct answer among the original and three collected endings.

difference between plausible and implausible answers is expected to be salient enough that it can be established as a classification task. However, when making day-to-day decisions, people consider several plausible choices, rather than clearly plausible and implausible ones, depending on one's situation and method of thinking. However, the tasks concerning conditions *under multiple plausible scenarios* are few, and their domains are limited to, for example, factual information that differs according to place and time (Zhang and Choi, 2021) or human behaviors that are either normative or divergent (Emelin et al., 2021). Another example is the natural language inference or commonsense reasoning task that considers variations in human

opinions (Zhang et al., 2017; Chen et al., 2020b), which allows for the differences in annotations due to one’s mentality (Pavlick and Kwiatkowski, 2019; Meissner et al., 2021). Our aim here is to interrogate these types of situated reasoning in more comprehensive settings, such as in story texts.

To assess the possible extent of situated reasoning in machines, we introduce *Possible Stories*, a benchmark consisting of 4,533 multiple-choice questions over 1,313 passages in English, to evaluate commonsense reasoning over multiple possible alternatives for single passages. Figure 1 shows an example. Given a story text, we aggregate alternative endings and multiple-choice questions that contain information such that they guide the determination of the most likely ending. By design, machines cannot rely only on answer options but also have to understand the condition implied by each question to answer correctly because all options are expected to be possible. This dataset creation procedure tackles the known issue of annotation artifacts (Gururangan et al., 2018) in crowdsourced datasets by using alternative endings, instead of *right* and *wrong* endings, and by compiling the endings and questions from multiple crowdworkers.

We evaluate strong pretrained language models and heuristic methods on our dataset and observe that in an unsupervised setting, even the strongest model (DeBERTa large v3; He et al., 2021) underperforms compared to humans by approximately 30% accuracy and more than 50% consistency score (i.e., passage-wise accuracy). Our analysis using input ablation and statistical significance tests highlights that the annotation artifacts contained in the answer options of our dataset questions are much fewer than those in existing multiple-choice datasets such as RACE (Lai et al., 2017) and CosmosQA (Huang et al., 2019). Reasoning-type annotation shows that more than 60% of our dataset questions require counterfactual reasoning, as well as an understanding of characters’ motivations and reactions, readers’ perceptions, and fictional information.

Our contributions are summarized as follows:¹

- We propose a situated commonsense reasoning task and create a multiple-choice question answering (QA) dataset using plausible story endings, together with questions as multiple condi-

tions where one of the endings becomes the most plausible.

- We discover that current strong pretrained language models struggle to solve our task when training data are unavailable, indicating that there is room for future improvement on situated commonsense reasoning.
- We show that our dataset contains minimal annotation artifacts in the answer options and has many challenging questions that require counterfactual reasoning and an understanding of characters’ motivations and reactions, readers’ perceptions, and fictional information.

2 Background and Related Works

Our work is motivated by recent efforts to create evaluation frameworks for commonsense reasoning situated in extra-linguistic contexts.

Benchmarks for Commonsense Reasoning

Many commonsense reasoning resources have been proposed that target reading comprehension (Huang et al., 2019), cloze tests regarding story endings (Zellers et al., 2019) or in-between events (Bhagavatula et al., 2020), and inferences on social interactions (Sap et al., 2019). Mostafazadeh et al. (2016) propose a task similar to ours, but it differs in that ours has four possible ending options, rather than a plausible and implausible completion pair.

Benchmarks for Counterfactual Reasoning

Researchers have coined the term *counterfactual reasoning* to refer to the property of reasoning over hypothetical events and have proposed benchmarks to evaluate the counterfactual reasoning ability of machines. Tandon et al. (2019) collect questions that explicitly ask *what if*, based on procedural texts. Qin et al. (2019) propose a task of generating a counterfactual story ending that is minimally edited from the original ending, given modified events in the context. Our data creation process is similar in terms of using an existing story and modifying a segment; however, we ask crowdworkers to change the segment more drastically, yielding diverse story endings.

Evaluation of Understanding of Situations

Reasoning over multiple possibilities, depending on the situation, can be regarded as *situated reasoning*. Recent studies have attempted to integrate situational information into the context used in downstream tasks, such as question answering on factual

¹The details of our data collection and final outcome including all collected story endings are available at <https://github.com/nii-cl/possible-stories>.

information (Zhang and Choi, 2021) and consequence or normative action generation given real-world social settings (Emelin et al., 2021). Story Commonsense (Rashkin et al., 2018) provides an annotated dataset of motivation and emotional reactions. (Forbes et al., 2020) collect general rules of thumb about actions. The range of situations that we consider goes beyond facts and normative settings, aiming to consider readers’ beliefs, causality, and characters’ emotions.

Probing of Language Models The use of contrastive examples to probe language models’ knowledge and inductive biases is an active area of research. This line of research typically uses pairs of sentences with minimum differences (Marvin and Linzen, 2018; Li et al., 2020; Warstadt et al., 2020), contrastive sets to identify the model’s decision boundary (Gardner et al., 2020), or adversarial examples (Jia and Liang, 2017) to identify the segments that contribute to changing model behaviors. By contrast, we use multiple plausible choices for a single passage to study what causes models to assign higher probabilities to certain choices.

3 Task Description

Motivation In Story Cloze Test, Mostafazadeh et al. (2016) use *right* and *wrong* endings to evaluate machines’ story understanding, assuming that the right ending can be regarded as an entailing hypothesis in a textual entailment framework and the wrong ending as a contradicting hypothesis. During data collection, the workers are instructed to produce endings that are realistic and sensible for right endings, and wrong endings are chosen from those that are rated lower than right endings in terms of meaningfulness and coherence. Consequently, their task is created to have clear right and wrong endings. However, in reality, there is an infinite number of possibilities of clearly plausible endings. By creating informative questions positing situations rather than questions asking about relative plausibility without any conditions, we aim to test machines’ story understanding in multiple scenarios that provide additional information that can discriminate one plausible ending from other possible endings.

Task Formulation We formulate the task as a multiple-choice question with a passage and answer options, where the answer options depict possible endings of the passage. Given a story s , the

task is to determine the most plausible story ending among the four possible endings $E = (e_1, \dots, e_4)$ under the condition c that is implied by a question. To further evaluate the models’ understanding of situations, we also define the task of predicting the most plausible outcome for multiple conditions. Given s , the task is to determine the most plausible story ending among E for each of the multiple conditions $C = (c_1, \dots, c_4)$ that are implied by multiple questions. When the models capture all the relationships between conditions and plausible endings correctly, we assume that the models reason over a finite number of possible consequences and the relationships among them. We call this *consistency*, which reports the percentage of a model’s outputs that are correct for all questions referring to a unique context. This evaluation is inspired by the study of contrastive sets (Gardner et al., 2020).

4 The Possible Stories Dataset

Context Passages To collect story texts, we use ROCStories (Mostafazadeh et al., 2016), a corpus of five-sentence stories. The first to fourth sentences describe the context, and the final sentence, the ending, concludes the story. We choose ROCStories because each of its stories has a clear beginning and ending, while being generic enough to come up with different endings. The details on our story selection criteria are provided in Appendix A.

The following tasks are carried out by the crowdworkers in Amazon Mechanical Turk (MTurk) who perform above certain levels during our worker recruitment phase, which is designed to be fairer than the conventional qualifications used in MTurk. The details of the worker recruitment are provided in Appendix B. The instructions and task interface presented to the workers are also provided in Appendix I.

4.1 Writing Tasks

Ending Writing We first ask workers to create two alternative endings given a story with the original ending. The participants are encouraged to be as creative as possible so that possible yet unrealistic story endings can also be elicited. We collect six to eight alternative endings by asking three or four workers to produce two endings per passage.

Selection of Ending Options Having collected six to eight alternative endings, we need to decide which three options to use in our questions, in addition to the original ending.

Depending on how they are chosen, there may be differences in the difficulty of the generated questions. For example, if the endings are similar to each other, it will be difficult to create questions that have only one correct choice among four endings. Conversely, if the endings are completely different, the questions may be easier to create, but machines may rely solely on semantic similarity between passage and endings, without requiring commonsense reasoning.

To examine the relationship between question difficulty and the diversity of the chosen endings, we run a pilot task using ten randomly selected stories with six different sets of endings. The six sets are chosen based on the sum of cosine similarity calculated based on the embeddings (Reimers and Gurevych, 2019) of all the possible combinations of endings, ranging from the set of endings that are most similar to the most diverse set. Six sets are chosen such that the distance between the values of the sum of the cosine similarities of one set and another set is equal. Through a validation step to identify which sets of endings enable high-quality multiple-choice questions, we decide on the set that contains the second most diverse endings among the six sets upon consideration.

Question Writing As four distinct endings are gathered per passage, we ask the workers to write questions in which only one of the four endings is the correct answer. Because this task is more complex, we select participants via a qualification task, targeting those participants who maintain quality in the ending writing task. It is up to the workers to decide the correct option, considering the difference in difficulty in writing questions with certain story endings. Four questions are written per passage by two workers, two per worker, and the answers to each set of two questions are different to maintain the diversity of the correct answers. To ensure that the distribution of the dataset is natural (Bowman and Dahl, 2021; Kaushik et al., 2021) and that the questions fit the general purpose, we avoid collecting questions in an adversarial manner (Bartolo et al., 2020).

4.2 Data Validation

The goal of the validation task is to verify that there is one correct answer for each question, and that the questions do not contain any objectionable or personal content. Questions that do not meet these criteria are discarded. The detailed validation re-

sults and further quality control over the collection batches are reported in Appendix C.

Question-Answer Validation In this step, we ask workers to answer multiple-choice questions. The workers choose one of the four endings and four additional options (no answer, more than two possible answers, ill-formed questions, and others). Each question is validated by three workers, and we retain questions in which the majority vote is identical to the writer’s answer.

Content Validation During the validation task, we ask workers to indicate negative stereotypes or biased descriptions of certain social groups. We discard questions that the workers claim contain unfair descriptions. This process prevents the perpetuation of unethical opinions in downstream tasks when this dataset is used for training models. Some of the workers’ inputs are discussed in Ethical Considerations. We incentivize workers with a bonus of \$0.3 per completion of the free-text form.

4.3 Dataset Statistics

Our dataset, Possible Stories, has 8,885 alternative endings for 1,313 passages and 4,533 multiple-choice questions with the original ending and three alternative endings as answer options. Table 1 presents the basic statistics for the resulting dataset. Although the passages are shorter than those in CosmosQA (70.3) and RACE (321.9), the questions (14.2) and answer options (15.3) are quite longer than others (CosmosQA is 10.6 and 8.1 and RACE is 10.0 and 5.3), which could potentially make questions difficult (Nangia et al., 2021).

In addition, as shown in Table 2, more than 50% of the contexts have questions with three or four distinct correct answer choices. This contributes to the assessment of the models’ comprehension of multiple situations using the consistency metric.

One of our main goals for constructing a benchmark is to test the models’ capacity for situated commonsense reasoning over multiple scenarios as an unseen task. Nonetheless, to ensure that it is feasible to model our task using current strong pretrained language models (Liu et al., 2019a), we follow a standard approach to split the collected examples into training (75%), dev (10%), and test (15%) sets. To investigate the model generalizability, the passages do not overlap between each set. The dev and test sets do not contain questions produced by workers who have received negative comments from other workers to ensure quality.

Split	#Question	#Passage	#Q/P	Passage len	Question len	Option len
Train	3,404	984	3.46	46.1	13.9	15.4
Dev	458	133	3.44	46.2	14.9	15.3
Test	671	196	3.42	47.0	15.0	15.2
Total	4,533	1,313	3.45	46.3	14.2	15.3

Table 1: Statistics of Possible Stories. Q and P indicate question and passage. $\#Q/P$ indicates the average number of questions per passage. Len indicates the average number of tokens.

#Q/P	Distinct # of answers				
	1	2	3	4	total
1	2.1				2.1
2	1.5	8.4			9.9
3		16.8	12.0		28.8
4		18.1	34.7	6.5	59.3
Total	3.6	43.3	46.7	6.5	100.0

Table 2: Distribution (%) of the number of questions per passage and the distinct number of correct answers.

5 Experiments

5.1 Models and Settings

For modern pretrained language models, we use BERT (base and large; Devlin et al., 2019), RoBERTa (base and large; Liu et al., 2019b), and DeBERTa (base and large of v3; He et al., 2021). In our standard setting (i.e., unsupervised), to adapt these models to the multiple-choice task, we fine-tune them on the RACE dataset (Lai et al., 2017), which is a large-scale dataset of middle- and high-school English exams and has passages and questions on various topics.² In the supervised setting, the models are directly trained on our training set unless mentioned otherwise. To establish different baseline methods, we consider simple heuristics using perplexity, semantic similarity, and entailment scores. For perplexity heuristics, we use GPT-2 (Radford et al., 2019) and GPT-Neo (Black et al., 2021) to obtain the perplexity of the inputs and consider options with the smallest perplexity as a model’s prediction. Sentence similarity uses representations obtained from the sentence transformers (Reimers and Gurevych, 2019) to compute the cosine similarity between the options and the rest of the input. The candidate with the highest similarity score is regarded as the model prediction. The en-

²We observe that models fine-tuned on CosmosQA are consistently inferior to those fine-tuned on RACE (Appendix D).

tailment score is calculated using RoBERTa-large fine-tuned on MNLI (Williams et al., 2018), and the option with the highest entailment score when taking the inputs as the premise is chosen.

5.2 Results

Human Performance To measure the human performance on our test set, we collect three additional labels for all questions from different crowdworkers who do not join the validation task. We ensure that the same set of three workers answer the questions belonging to a single story. For computing accuracy, we take the majority of the three labels to determine whether it is equal to the validated gold label. For computing consistency, we determine whether the majority vote answers are correct for all questions in each passage (Table 3).

Model Performance When the training set is unavailable, we observe that DeBERTa-large achieves the best performance. Although this model is fine-tuned on RACE, which has a sufficient number of diverse training examples, the model performance is far from that of humans, showing large gaps of 29.5% and 53.0% in terms of accuracy and consistency, respectively. Out of the four simple heuristics models, those using perplexity and semantic similarity perform above the chance rate of 25%, indicating that those features, while inadequate, might be useful in finding the correct answers. By contrast, the entailment score-based model falls short of 25%. This result highlights the uniqueness of our dataset, as it shows that relying on monotonic reasoning cannot lead to a correct answer.

Supervised Performance With the training data, we observe that DeBERTa-large performs better than the other models, and it achieves the best accuracy and consistency when fine-tuned using RACE (Table 3). These scores are very close to those of humans, which implies that the task can be feasibly performed by a strong model, given sufficient train-

FT	Model	Acc.	Consist.
	DeBERTa-large*	60.2	19.9
	DeBERTa-base*	45.3	8.2
	RoBERTa-large*	50.5	13.8
✗	PPL. GPT-2 large	30.4	2.0
	PPL. GPT-Neo 2.7B	29.5	2.6
	Semantic Sim.	37.3	4.1
	Entailment	23.1	2.0
	DeBERTa-large*	92.1	74.7
	DeBERTa-large	88.5	67.3
	DeBERTa-base	81.5	51.5
✓	RoBERTa-large*	83.5	55.6
	RoBERTa-large	81.7	49.5
	RoBERTa-base	72.0	30.6
	BERT-large	62.6	20.4
	BERT-base	57.3	16.3
	Human	92.5	76.5

Table 3: Model and human performances on our dataset. *Acc.* and *consist.* denote accuracy (%) and consistency (%). (*) indicates that the model is fine-tuned on RACE. *FT* indicates whether the models are fine-tuned on the training set. The experimental details are reported in Appendix E.

ing data. Nonetheless, it is notable that BERT-large and RoBERTa-large, which were state-of-the-art models only several years ago, show potential for improvement ($\approx 30\%$ and 10% accuracy, respectively) compared to humans.

Input Ablation Table 4 presents the input ablation analysis. When ablating the passages, we observe that pretrained language models fine-tuned on any multiple-choice dataset show lower performance than those with the full input. Regarding the heuristics methods, we find that having the context does not significantly change the ranking of the endings. When ablating the questions, we observe that the performance of all models decreases, which is expected because the same set of answer options has multiple questions in our dataset.

6 Analysis

6.1 Human–model Performance Gap

To investigate the relative difficulty of our dataset among multiple-choice QA datasets, we compare the accuracy gap between humans and models in an unsupervised setting with existing datasets, including CosmosQA (Cosmos; we report the valida-

FT	Model	Full	No pas.	No ques.
	DeBERTa-L*	60.2	58.1	21.8
✗	RoBERTa-L*	50.5	50.3	21.5
	GPT-2 large	30.4	35.2	26.4
	Semantic Sim.	37.3	47.1	28.8
	DeBERTa-L*	92.1	87.0	31.8
✓	DeBERTa-L	88.5	86.4	33.4
	BERT-L	62.6	51.1	30.4

Table 4: Input ablation results (accuracy; %). *No pas.* and *no ques.* indicate that the context passage and question are ablated from the input, respectively.

Model	Ours	Cosmos	QuAIL	MC-adv
DeBERTa-L	60.2	66.8	76.3	81.2
DeBERTa-B	45.3	56.0	66.2	69.0
RoBERTa-L	50.5	64.2	70.3	69.1
Human	92.5	94.0	100.0	92.0
Acc. gap	40.5	31.7	29.1	18.9

Table 5: Human performance, model performance without fine-tuning (accuracy; %), and the gap between the human performance and the average model performance (larger values imply higher difficulty). Model-L and -B indicate the large and base models respectively.

tion result because the test labels are not available), QuAIL (Rogers et al., 2020, challenge set), and the examples provided by Sugawara et al. (2022) (MC-adv; multiple-choice questions that are written by crowdworkers in an adversarial manner). We use three models (DeBERTa-large and -base and RoBERTa-large) fine-tuned on RACE. The results in Table 5 demonstrate that our dataset may be more challenging than the multiple-choice reading comprehension datasets we analyze, despite the simplicity of our data collection method.

6.2 Annotation Artifacts in Answer Options

One of our motivations for crowdsourcing multiple questions for the same set of answer options is to minimize superficial patterns (i.e., annotation artifacts) in the collected examples, especially in their answer options. To validate this, we first compare the supervised performance in three ablation settings (no context passage, no question, and answer options only). We use DeBERTa-large and report the test score on our dataset, RACE, QuAIL, and CosmosQA. Table 6 shows that although the no-context performance on our dataset

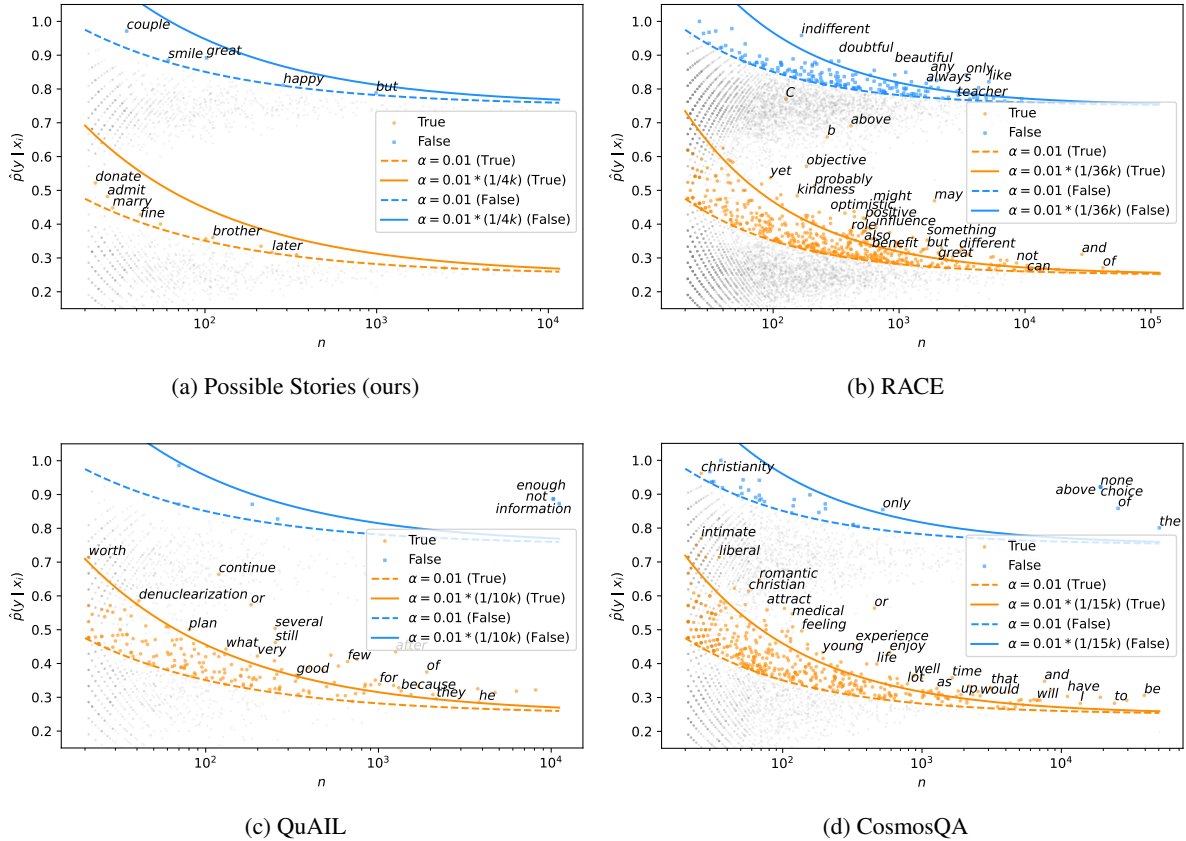


Figure 2: Token-level annotation artifacts in the training examples of our dataset, RACE, QuAIL, and CosmosQA. All tokens are below $\alpha = 0.01$ with a conservative Bonferroni correction for 3,990, 15,472, 35,762, and 9,688 vocabulary items, respectively.

Dataset	Full	No pas.	No ques.	Opt. only
Ours	88.5	86.4	33.4	29.1
RACE	87.9	60.1	69.8	46.6
QuAIL	81.7	51.8	58.3	39.6
Cosmos	87.8	72.5	59.4	57.2

Table 6: Supervised accuracy (%) by DeBERTa-large (v3) on the input-ablation settings.

is relatively high, the no-question and option-only performances are lower than the others. This result implies that the question and answer options in our dataset are mutually indispensable for predicting the correct answer, while in the other datasets, the options on their own and their relationship with the context are informative for the prediction.

To visualize the actual tokens that create annotation artifacts, we follow Gardner et al. (2021), who propose analyzing token-level features in terms of the empirical probability of labels $\hat{p}(y|x_i)$ given a specific token (vocabulary item) x_i appearing in input X . Here, the label y indicates whether an

answer option is the correct (True) or not (False). We plot the probability $\hat{p}(y|x_i)$ and the number of occurrences (n) for the tokens of the training questions in our dataset, RACE, QuAIL, and CosmosQA (Figures 2) for comparison. To see if the null-hypothesis (i.e., the token does not co-occur with a specific label) is accepted or rejected, we compute z -statistics and plot the level of statistical significance $\alpha = 0.01$ and its conservative Bonferroni correction (Bonferroni, 1936) for the vocabulary items ($\alpha = 0.01/|V|$). We find that for the true label in our dataset, only 12 tokens are above $\alpha = 0.01$ and no tokens are above $\alpha = 0.01/|V|$ where several content words, such as *admit*, *fine*, and *great* are possibly helpful for predicting the correct label. By contrast, 421 and 84 tokens in CosmosQA are found to be statistically significant at the respective levels, where many content words, such as *enjoy* and *life*, function words, such as *or*, and the task-specific phrase (*none of the above choice*) are strong indicators. We observe similar trends for RACE and QuAIL. In Appendix F, we report the numbers of vocabulary items above the



Figure 3: Question words and the subsequent words of the questions in our dataset.

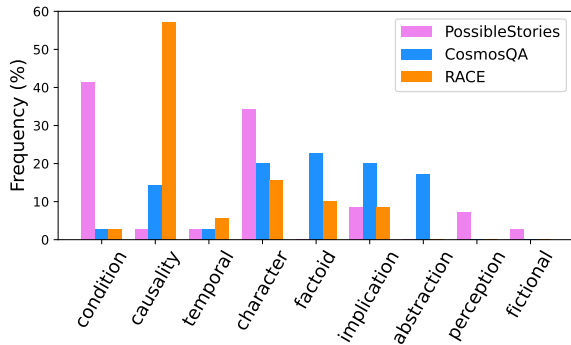


Figure 4: Reasoning types across our dataset, CosmosQA, and RACE.

levels of statistical significance for the four datasets. These results show that the answer options in our dataset contain minimal annotation artifacts compared to those of the other datasets.

6.3 Question and Reasoning Types

The question words and subsequent words in the test questions are plotted in Figure 3. We find that more than half the number of questions are *what* questions, seemingly asking about the concrete content of the story. We also observe subsequent words, such as *would*, *outcome*, and *if*, which lead to a statement requiring commonsense reasoning.

To investigate the kind of reasoning required for answering, we annotate the collected questions with reasoning types. Considering previous studies, we define nine reasoning types (See Appendices G and H for the definitions and examples). We annotate 70 questions from our test set and the same number of questions taken from CosmosQA and

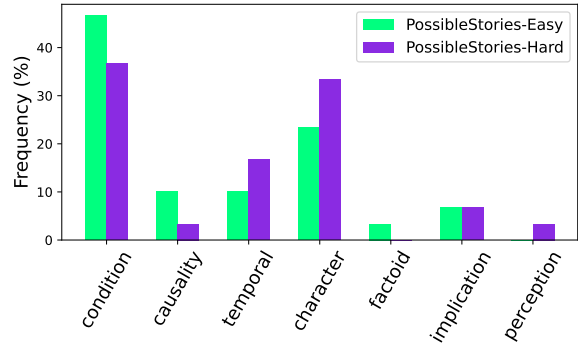


Figure 5: Reasoning types of easy and hard questions in our dataset.

RACE for comparison (Figure 4). In addition, to examine the relationship between question difficulty and reasoning types, we split the test examples into easy and hard subsets and annotate 30 questions for each subset (Figure 5). The easy questions are those for which the human-model accuracy gap is 0% in terms of accuracy, and the hard ones are those for which the gap is larger than 60% (215 and 64 examples). To compute model performance, we average the accuracy of the five models (BERT-base and large, RoBERTa-base and large, and RoBERTa-large, which is fine-tuned on RACE; all models are fine-tuned on our training set). Apart from the reasoning types, we independently check whether each question requires counterfactual reasoning. This includes not only the condition and fiction types but also other types such as temporal reasoning in the sense that it can be involved in reasoning over counterfactual conditions. The frequency (%) of such questions is as follows: ours 68.6, ours-hard 76.7, ours-easy 66.7, CosmosQA 4.3, and RACE 2.9.

In summary, our major findings are as follows:

- Our dataset includes more questions regarding conditions and characters’ motivations and reactions than the other datasets. It also has a small number of fictional and perception questions, while the others do not.
- We do not observe abstraction and factoid questions in the annotated examples. However, we find several abstraction questions in our test set, one of which is presented in Appendix H.
- Questions regarding characters’ motivations and reactions are relatively harder, while questions regarding causality, which do not require counterfactual reasoning in most cases in our annota-

P: The Smith family loved to go on day trips on their boat in the summer. One day, they decided it would be fun to take the kids to a new place. They chose to travel north to a beach that wasn't terribly far away. The children had a wonderful time and met a new friend to play with.

Q1: Which of these is the most negative ending? **Q2:** Which of these implies that the trip they took was successful?
Q3: Which ending implies the Smith kids were bad at staying in touch? **Q4:** Which ending involves the most conflict?

Q1 Q2 Q3 Q4 Options

- | | | | | |
|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|---|
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | A: They kept in touch with their friend even after they went home. |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input checked="" type="checkbox"/> | B: At the end of the day the kids got into a fight with each other and were happy to leave. |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | C: The Smith's decided they'd visit a new beach every year, and they made tons of new friends. |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> | D: They went home though and the kids never saw their friend again. |

Figure 6: Example of questions with a single passage. Check mark (✓) indicates the correct option. Cross mark (✗) indicates that DeBERTa-large (v3) makes an incorrect prediction with that option.

tion, are easier. This corresponds with the fact that we find more counterfactual questions in the hard questions than in the easy questions.

6.4 Case Study

We present examples in which the strongest model (DeBERTa-large) makes incorrect predictions (Figure 6). A single worker writes Q1 and Q2 and another worker writes Q3 and Q4. Q1 and Q4 are annotated as perceptions (*the most negative ending* and *the most conflict*). It seems that the model struggles to compare which option is more negative between options B (*got into a fight...were happy*) and D (*never saw their friend*). Q2 and Q3 are annotated as implications (*imply...*). For Q2, we must infer that option C (e.g., *made tons of new friends*) implies success, but option B (*kept in touch with their friend*) might sound more successful to the model. More examples of other reasoning types are provided in Appendix H.

7 Discussion

Circumscribing commonsense reasoning from simple heuristics has been a long-standing problem in the field of artificial intelligence (Levesque, 2014). Although the answer options in our dataset are free from annotation artifacts, our ablation analysis in Section 5 also shows that the questions and answer options may still involve some clues that the models can exploit. Further research is needed to explain how commonsense reasoning is distinguished from a set of simple heuristics in machines' situational understanding.

In the question-writing task, one of the workers addresses the difficulty of creating questions that cannot be answered without reading the passages and that it is even practically impossible unless asking the questions to small children. This illustrates

that humans may also use a small amount of information available to draw inferences. This issue arises possibly because of our task formulation, answering which option is more plausible than the others. It can be argued that modifying this task to a generative task (Chen et al., 2020a) is one way to directly assess machines' commonsense reasoning ability, but it should be noted that this would entail some difficulties in the evaluation of generated answers.

In addition, exploring what kind of conditions narrow down the possibilities of consequences is important to effectively evaluate machines' situational understanding. Although several studies have captured the dynamics of conditions in moral and immoral settings (Emelin et al., 2021), many other factors come into play in decision-making in our daily lives, such as feelings, personal beliefs, expectations from others, or even unconscious biases.

8 Conclusion

This paper proposes a new dataset, Possible Stories, consisting of 4.5K crowdsourced questions with 1.3K story passages to investigate whether machines can infer the most plausible ending among four possible endings under certain situations postulated by questions. We discover that current strong pretrained language models struggle to answer questions consistently, showing a large accuracy gap compared with humans. A comparison with existing multiple-choice datasets demonstrates that our questions contain minimal annotation artifacts in the answer options and require counterfactual reasoning as well as an understanding of characters' motivations and emotions, suggesting that our dataset can serve as a challenging benchmark for future commonsense reasoning studies.

Ethical Considerations

This study aims to facilitate the scientific study of machines’ situated commonsense reasoning. We use crowdsourcing for our data collection, taking care to avoid the exploitation of workers and pay well above the U.S. federal minimum wage. The details of worker recruitment and the payment process are described in Appendix B. We also validate that the examples in our dataset do not contain unfair or harmful content. In this section, we report our observations regarding the validation task. This study is approved by our internal review board.

Content Validation for Fair Representation

During content validation (Section 4.2), we find that the level of content to be flagged is not trivial. There is a question containing the phrase *mainstream COVID-related propaganda*, and one of the workers told us that the worker was unsure if it should be flagged. Another case involves a story ending that describes the cooking skills of a male character in a bad light. Does this representation perpetuate the negative stereotype that men are bad at cooking? To investigate this, we should dive deeper into the semantic plausibility learned in language models (Porada et al., 2021; Pedinotti et al., 2021). Unless the focus is on the domain of natural science, there is less agreement on what would lean in spreading desirable and undesirable content, and the borderline can change across time and place. It should also be noted that the degree of sensitivity towards underspecified biases depends on individuals’ imagination and empathy. Future work can examine how to effectively moderate the dataset for fair and unbiased representation.

Limitations One of the limitations of the study is “limited diversity.” We observe that some systemic biases during data collection. One example concerns a story in which the protagonist missed breakfast on a day of work. Many crowdworkers come up with the possibility of the girlfriend bringing the lunch to the protagonist’s workplace (referred to as *I* throughout the context), but no one assumes that the boyfriend will do the same. These types of unconscious biases can accumulate in datasets. In addition, our dataset is limited to English.

Acknowledgments

The authors would like to thank the anonymous reviewers for their insightful comments. This work

was supported by JST PRESTO Grant Number JP-MJPR20C4 and JSPS KAKENHI Grant Number 22K17954.

References

- Shourya Aggarwal, Divyanshu Mandowara, Vishwa-jeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. 2021. [Explanations for CommonsenseQA: New Dataset and Models](#). 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 3050–3065, Online. Association for Computational Linguistics.
- Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. [Beat the AI: Investigating adversarial human annotation for reading comprehension](#). *Transactions of the Association for Computational Linguistics*, 8:662–678.
- Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen tau Yih, and Yejin Choi. 2020. [Abductive commonsense reasoning](#). In *International Conference on Learning Representations*.
- Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. [GPT-Neo: Large scale autoregressive language modeling with mesh-tensorflow](#).
- Carlo E. Bonferroni. 1936. [Teoria statistica delle classi e calcolo delle probabilita](#). *Pubblicazioni del R. Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62.
- Samuel R. Bowman and George Dahl. 2021. [What will it take to fix benchmarking in natural language understanding?](#) In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online. Association for Computational Linguistics.
- Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2020a. [MOCHA: A dataset for training and evaluating generative reading comprehension metrics](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6521–6532, Online. Association for Computational Linguistics.
- Tongfei Chen, Zhengping Jiang, Adam Poliak, Keisuke Sakaguchi, and Benjamin Van Durme. 2020b. [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.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of](#)

- deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. 2021. [Moral stories: Situated reasoning about norms, intents, actions, and their consequences](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 698–718, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. [Social chemistry 101: Learning to reason about social and moral norms](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.
- Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. [Evaluating models’ local decision boundaries via contrast sets](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.
- Matt Gardner, William Merrill, Jesse Dodge, Matthew Peters, Alexis Ross, Sameer Singh, and Noah A. Smith. 2021. [Competency problems: On finding and removing artifacts in language data](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1801–1813, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- 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.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. [DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing](#). arXiv preprint 2111.09543.
- Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. [Cosmos QA: Machine reading comprehension with contextual commonsense reasoning](#). 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 2391–2401, Hong Kong, China. Association for Computational Linguistics.
- Robin Jia and Percy Liang. 2017. [Adversarial examples for evaluating reading comprehension systems](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.
- Divyansh Kaushik, Douwe Kiela, Zachary C. Lipton, and Wen-tau Yih. 2021. [On the efficacy of adversarial data collection for question answering: Results from a large-scale randomized study](#). 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 6618–6633, Online. Association for Computational Linguistics.
- Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. [UNIFIEDQA: Crossing format boundaries with a single QA system](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online. Association for Computational Linguistics.
- Jonathan K. Kummerfeld. 2021. [Quantifying and avoiding unfair qualification labour in crowdsourcing](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 343–349, Online. Association for Computational Linguistics.
- Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. [RACE: Large-scale Reading comprehension dataset from examinations](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.
- Hector J. Levesque. 2014. [On our best behaviour](#). *Artificial Intelligence*, 212:27–35.
- Tao Li, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Vivek Srikumar. 2020. [UNQOVERing stereotyping biases via underspecified questions](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3475–3489, Online. Association for Computational Linguistics.
- Nelson F. Liu, Roy Schwartz, and Noah A. Smith. 2019a. [Inoculation by fine-tuning: A method for analyzing challenge datasets](#). 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 2171–2179, Minneapolis, Minnesota. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. [RoBERTa: A robustly optimized BERT pretraining approach](#). arXiv preprint 1907.11692.
- Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. [UNICORN on RAINBOW: A universal commonsense reasoning model on a new multitask benchmark](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13480–13488.
- Rebecca Marvin and Tal Linzen. 2018. [Targeted syntactic evaluation of language models](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.
- Johannes Mario Meissner, Napat Thumwanit, Saku Sugawara, and Akiko Aizawa. 2021. [Embracing ambiguity: Shifting the training target of NLI models](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 862–869, Online. Association for Computational Linguistics.
- Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. [A corpus and cloze evaluation for deeper understanding of commonsense stories](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.
- Nasrin Mostafazadeh, Aditya Kalyanpur, Lori Moon, David Buchanan, Lauren Berkowitz, Or Biran, and Jennifer Chu-Carroll. 2020. [GLUCOSE: Generalized and Contextualized story explanations](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4569–4586, Online. Association for Computational Linguistics.
- Nikita Nangia, Saku Sugawara, Harsh Trivedi, Alex Warstadt, Clara Vania, and Samuel R. Bowman. 2021. [What ingredients make for an effective crowdsourcing protocol for difficult NLU data collection tasks?](#) 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 1221–1235, Online. Association for Computational Linguistics.
- Ellie Pavlick and Tom Kwiatkowski. 2019. [Inherent disagreements in human textual inferences](#). *Transactions of the Association for Computational Linguistics*, 7:677–694.
- Paolo Pedinotti, Giulia Rambelli, Emmanuele Chersoni, Enrico Santus, Alessandro Lenci, and Philippe Blache. 2021. [Did the cat drink the coffee? challenging transformers with generalized event knowledge](#). In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 1–11, Online. Association for Computational Linguistics.
- Ian Porada, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. 2021. [Modeling event plausibility with consistent conceptual abstraction](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1732–1743, Online. Association for Computational Linguistics.
- Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. 2019. [Counterfactual story reasoning and generation](#). 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 5043–5053, Hong Kong, China. Association for Computational Linguistics.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. [Language models are unsupervised multitask learners](#). *OpenAI blog*, 1(8):9.
- Hannah Rashkin, Antoine Bosselut, Maarten Sap, Kevin Knight, and Yejin Choi. 2018. [Modeling naive psychology of characters in simple commonsense stories](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2289–2299, Melbourne, Australia. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. [SentenceBERT: Sentence embeddings using Siamese BERT-networks](#). 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. 2020. [Getting closer to AI complete question answering: A set of prerequisite real tasks](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8722–8731.
- Swarnadeep Saha, Prateek Yadav, Lisa Bauer, and Mohit Bansal. 2021. [ExplaGraphs: An explanation graph generation task for structured commonsense reasoning](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*,

- pages 7716–7740, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. [WinoGrande: An adversarial Winograd schema challenge at scale](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8732–8740.
- 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.
- Saku Sugawara, Nikita Nangia, Alex Warstadt, and Samuel Bowman. 2022. [What makes reading comprehension questions difficult?](#) In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6951–6971, Dublin, Ireland. 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.
- Niket Tandon, Bhavana Dalvi, Keisuke Sakaguchi, Peter Clark, and Antoine Bosselut. 2019. [WIQA: A dataset for “what if...” reasoning over procedural text](#). 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 6076–6085, Hong Kong, China. Association for Computational Linguistics.
- Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. [BLiMP: The benchmark of linguistic minimal pairs for English](#). *Transactions of the Association for Computational Linguistics*, 8:377–392.
- Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. [A broad-coverage challenge corpus for sentence understanding through inference](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. [HellaSwag: Can a machine really finish your sentence?](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.
- Michael Zhang and Eunsol Choi. 2021. [SituatQA: Incorporating extra-linguistic contexts into QA](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7371–7387, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- 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 Selecting Stories from ROCStories

We choose stories with more than 45 words and endings with more than 5 words to avoid stories that are too short or generic. We consult the annotation of the GLUCOSE dataset (Mostafazadeh et al., 2020) and select stories whose ending is annotated by workers with ratings higher than 1. Future work could investigate the effect of different causal relations on creating conditions, as well as possible endings.

B Crowdsourcing Recruitment and Payment

We recruit writers via Amazon Mechanical Turk (MTurk). The number of workers who participated in the study is listed in Table 7. Before initiating the data collection procedure, we first run a qualification task to identify workers who can participate in data collection. This task is a short version of a part of the main task and is open to any crowdworker without any qualifications such as HIT acceptance rate or number of HIT accepted, which are commonly used thresholds in the NLP community’s data collection process via MTurk.

	Ending	Ques.	Valid.
# of crowdworkers	163	66	65
Ave. # of examples	54.6	79.9	243.3
Max. # of examples	132	170	400

Table 7: Statistics of crowdworkers that participate in each task, the average and maximum number of generated examples per crowdworker.

We adapt this qualification following the recommendation of Kummerfeld (2021) to avoid the exploitation of crowdworkers. He demonstrates that imposing these prepared criteria is not fair because crowdworkers need to work on poorly paid tasks to achieve those qualifications in most cases.

We pay \$1.0 USD for an ending writing task, \$1.5 for a question writing task, and \$1.0 for a validation task, estimating the completion time to be less than 5, 7.5, and 5 mins respectively. This adds to more than \$12, which is well above the U.S. federal minimum wage. We do not calculate the wage according to the cost of living in each country where the workers reside, as we do not ask them where they live.

C Validation Results and Quality Control

Validation Results During the question-answer validation, 13.8% of the collected questions are discarded. Out of the four additional options, questions with no answer account for 1.8%, those with more than two possible answers account for 6.8%, ill-formed questions account for 1.8%, and others account for 2.1% of the total. The high frequency of questions with more than two options is understood to be due to the possibility that some answer options are too similar to each other to create questions with a single correct answer. Through the content validation process, 0.2% of the questions are discarded.

Quality Control During the data collection process, we repeat all tasks three times (i.e., three batches). The first and second batches have no workers in common, resulting in 52% of the final dataset with a total of 66 workers. For the final batch, we further qualify the workers who participated in these batches using three criteria: 1) writing more than nine questions, 2) mean human validation accuracy of more than 66%, and 3) creating more than 90% of questions *aswh*-questions to ensure dataset quality. Additionally, we manu-

Model	RACE	CosmosQA
DeBERTa-large [†]	92.1	89.7
DeBERTa-large	88.5	51.3
RoBERTa-large [†]	83.5	83.3
RoBERTa-large	50.5	38.3

Table 8: Accuracy (%) of models on our test set that are fine-tuned on RACE and CosmosQA respectively. [†] indicates that the model is trained on our training set (i.e., supervised).

Model	b	lr
DeBERTa-large	24	1e-5
DeBERTa-base	48	3e-5
RoBERTa-large	24	1e-5
RoBERTa-base	48	3e-5
BERT-large	36	1e-5
BERT-base	72	3e-5

Table 9: Hyperparameters used in the experiments. b and lr indicate the batch size and learning rate, respectively.

ally check the comments given to each worker and exclude workers who tend to produce yes/no questions and those containing unethical or politically sensitive topics. The final batch yields 48% of the final dataset with 38 workers.

D Comparison of Models Fine-tuned on RACE and CosmosQA

In our experiments, we use RACE for fine-tuning our pretrained language models to adapt them to the multiple-choice task. This is because we observe that RoBERTa-large and DeBERTa-large fine-tuned on RACE show higher performance than the corresponding models fine-tuned on CosmosQA (Table 8) in both unsupervised and supervised settings.

E Details of Experiments

Table 9 reports the hyperparameters used in our experiments. We use Huggingface’s Transformers library (Wolf et al., 2020) for our experiments.

Table 10 reports the detailed results of DeBERTa-large (fine-tuned on RACE) on our test set in the unsupervised and supervised settings. Owing to computational constraints, we conduct five different runs only for this model, which is the strongest among the models we use in our experiments. We

FT	Model	Full input		No passage		No question	
		Accuracy	Consist.	Accuracy	Consist.	Accuracy	Consist.
✗	DeBERTa-large*	60.2±1.7	19.9±2.2	58.1±2.6	19.9±1.7	21.8±1.6	0.5±0.4
✓	DeBERTa-large*	92.1±0.6	74.7±2.3	87.0±0.7	62.1±1.8	31.8±1.6	1.9±0.7

Table 10: Unsupervised and supervised performance (%) with the standard deviations of DeBERTa-large in five runs. The five models are fine-tuned on RACE with different random seeds, respectively.

α	Ours	Cosmos	RACE	QuAIL
0.01	12/5	421/33	475/163	173/7
0.01/ V	0/0	84/6	104/19	39/3
V	3,990	15,472	35,762	9,688

Table 11: The number of vocabulary items that appear in correct/incorrect options above the levels of statistical significance ($\alpha = 0.01$ and its conservative Bonferroni correction for the size of vocabulary $|V|$).

do not observe large deviations across the runs.

F Annotation Artifacts in Answer Options

We report the number of examples above different levels of statistical significance across the four analyzed datasets in Table 11. The number for our dataset above $\alpha = 0.01/|V|$ is zero, whereas those for the other datasets are significantly larger. This result shows that our dataset does not suffer from token-level annotation artifacts in the answer options, supporting our findings on the option-only training results in Section 6.1.

G Definitions of Reasoning Types

Annotating reasoning types is not a trivial task, particularly because the questions are fully written by humans without templates. Moreover, it is possible to use many classification methods, and there is rarely a consensus on reasoning types. For example, CosmosQA proposes seven reasoning types: pre/post-conditions, motivations, reactions, temporal events, situational facts, counterfactuals, and other (e.g., cultural norms). In QuAIL, nine types of reasoning are proposed spanning three categories: temporal, factoid, character properties for *text-based questions*, coreference, causality, belief states, subsequent entity states, event durations for *questions that require world knowledge*, and *unanswerable*. After categorizing these into five types, we add three types: abstraction (summarizing what hap-

1. **Condition:** pre/post counterfactual conditions introduced in the question.
2. **Causality:** causes and effects of events.
3. **Temporal:** temporal relations between events.
4. **Character:** characters’ emotions, motivations, and reactions.
5. **Factoid:** extracting entities from the context.
6. **Abstraction:** lesson, conclusion, and summary of the context.
7. **Implication:** paraphrasing and implication about events.
8. **Perception:** reader’s perceptual responses.
9. **Fictional:** fictional situations as counterfactual condition.

Table 12: Definitions of reasoning types.

pened), implication (paraphrasing), and readers’ (observers’) perceptions. In addition, we differentiate reasoning over fiction from counterfactual in that it is a more specific type of counterfactual that is considered implausible for most people in the real world. This results in the nine reasoning types listed in Table 12. Appendix H presents some examples.

H More Examples for Reasoning Types and Difficulty

The reasoning types and their example questions taken from our dataset are listed in Table 13. We also show examples of passage, question, and answer options in our dataset, including easy and hard questions, in Figure 7. Each question ends with its reasoning type and easy/hard classification, if available.

I Annotation Instructions and Interfaces

I.1 Ending Writing

Figures 8 and 9 show the instructions used in the story ending writing task. Figure 10 shows the interface used in the story ending writing task.

Reasoning	Example
Condition	Jeff is a child with a very vivid sense of imagination. What is most likely to have happened next?
Causality	Which is the most likely caused the guests to avoid shards of glass?
Temporal	Which is most likely if Chris later felt sick to his stomach?
Character	What outcome would be most upsetting to Ben?
Factoid	Where did people hide the money they got?
Abstraction	What lesson did she learn from the passage?
Implication	Which answer implies Bob was pleased with his performance?
Perception	What is the most moral decision for Danielle?
Fictional	How does Dylan get home?

Table 13: Reasoning types we use in the annotation and their example questions.

I.2 Question Writing

Figures 11, 12, and 13 show the instructions used in the question writing task. Figures 14 and 15 show the interface used in the question writing task.

I.3 Question Validation

Figure 16 shows the instructions used in the question validation task.

P1: Lydia was listening to an old CD her boyfriend had burned for her. Her CD player was old but still working alright. She had lost track of her thoughts and was enjoying the music. Suddenly, the CD skipped out and stopped playing.

Q1: Why was the CD player unable to function? (causality, easy)

Q2: Which answer indicates that Lydia would never be able to listen to the CD again? (implication, hard)

Q3: Which of the following is likely to occur if we know Lydia has realized the CD player cannot be fixed? (condition, hard)

Q1 Q2 Q3 Options

- A:** Lydia tried to fix it but the CD had a huge scratch.
- B:** Lydia tinkered with the CD player and got it working again.
- C:** Lydia went to bed upset, knowing she had to buy a new one in the morning.
- D:** She realized the batteries in her CD player had died.

P2: Darrel was waiting in the drive through for half an hour. He had about lost his patience. When he finally got to the window he was about to scream at them. They immediately apologized before he could.

Q1: How did the employees react when they saw Darrel's face turn red at the drive-through window? (character)

Q2: How did Darrel respond after the employees apologized for the long wait? (character, easy)

Q3: If Darrel's mind was soon preoccupied with something entirely different, what was most likely to have happened? (condition, easy)

Q4: In this scenario, what most likely happened if Darrel was pleased soon thereafter? (character, hard)

Q1 Q2 Q3 Q4 Options

- A:** They had an accident and offered free food to make it up to him.
- B:** He chose not to accept the apology and asked to speak to the manager.
- C:** They quickly gave him his food and informed him that there were very few employees working that day.
- D:** Before he could open his mouth, his engine started smoking and he had to call a tow truck.

P3: Jan checked to make sure no one was around. Her two older brothers had been sneaking around the garden lately. Being a curious child, Jan wanted to know what they were up to. She carefully opened the door to her brother's room.

Q1: If Jan smelled pleasant aromas and felt fresh air in the room, what did she likely discover? (condition)

Q2: What was the likely outcome if Jan was left still feeling clueless about what her brothers had been up to? (character)

Q3: Which outcome is the most unlikely to occur in reality? (fiction)

Q4: Which would be particularly unpleasant for Jan if she suffers from acute arachnophobia? (character)

Q1 Q2 Q3 Q4 Options

- A:** Inside the back of their closet, she found several jars with spiders.
- B:** There was a strange looking alien peeking out of a corner with fearful eyes.
- C:** They had taken plants from the garden and moved them to their room.
- D:** The door slammed shut on her face as the cameras alerted the brothers of an intruder.

P4: Billy liked Christmas songs. But didn't know what a turtle dove was. He like turtle and knew they were green and had a shell. He also knew what a dove was, a type of bird.

Q1: What happened if it was the worst Christmas of Billy's life? (condition, easy)

Q2: What happened if he pictured a turtle with wings? (fictional, easy)

Q3: What outcome would be most tragic? (perception)

Q1 Q2 Q3 Options

- A:** So he decided that 12 drummers drumming was a better part of the song.
- B:** He decided that a turtle dove was likely a flying turtle.
- C:** Billy became a famous author after embracing his love for holiday traditions.
- D:** He went to ask his mother about turtle doves, but when he found her in the bathtub, she was dead.

Figure 7: Examples in our dataset. Check mark (✓) indicates the correct option. Cross mark (✗) indicates that RoBERTa-large fine-tuned on RACE and our training set makes an incorrect prediction with that option.

Transform a story ending to make it unique.

Thank you for your participation. Please read the following instructions (and examples) carefully.

Instructions

You will be presented a short story.

Your task;

- Create **two** story endings (each in **ONE** sentence) relying on the one original passage (**Alternative Endings**). The two endings you wrote must be **totally different** from each other.

We recommend spending 5-10 minutes on this HIT.

Guidelines for the Task

The goal here is to modify stories to make them those *you* think are unique, less stereotypical, less usual, but those other people will still be able to enjoy it as a story.

Here are some guidelines:

- You are encouraged to be **as creative as possible**, such as;
 - imagining yourself as a different person and describing how that person would behave in the given condition.
 - introducing extraterrestrial beings, animals that can communicate with humans.
- Consider modifying the stories while paying attention to:
 - Characters' feelings and motivations
 - Causes and consequences of described events
 - Definition, properties, and process explained in a passage
 - Summary and lesson of a passage
- Use words and phrases that don't appear in the passage as many as possible.
- Write an ending in **ONE** sentence.
- Write story endings that are **quite different** from the original story endings.
- Also, the two story endings you wrote must be **quite different** from each other.
- Before submission, make sure you read the whole passage with the ending you write so that you can confirm that the story is possible.

Figure 8: Instructions (1/2) used in the story ending writing task.

Example

Tina made spaghetti for her boyfriend. It took a lot of work, but she was very proud. Her boyfriend ate the whole plate and said it was good. Tina tried it herself, and realized it was disgusting.

Original Ending: She was touched that he pretended it was good to spare her feelings.

Example Response #1

Tina made spaghetti for her boyfriend. It took a lot of work, but she was very proud. Her boyfriend ate the whole plate and said it was good. Tina tried it herself, and realized it was disgusting.

Alternative Ending: She got upset that he hadn't been honest with her.

Example Response #2

Tina made spaghetti for her boyfriend. It took a lot of work, but she was very proud. Her boyfriend ate the whole plate and said it was good. Tina tried it herself, and realized it was disgusting.

Alternative Ending: Soon after, she realized that she was dreaming.

IMPORTANT FINAL NOTE

- Please **AVOID** writing sentences that involve making prejudicial assumptions about people based on their identities.
- Please **AVOID** writing extremely violent endings, or using abusive language such as slurs.
- Please refer to this FAQ for additional information.

Figure 9: Instructions (2/2) used in the story ending writing task.

Given the passage, create two story endings (each in ONE sentence, totally different from each other) relying on the original passage (**Alternative Endings**). You can accept up to 25 HITs for this batch. **Please read the full instructions before starting.**

▼ Passage 1 / 1

In Progress → Completed

Passage

The delivery man was riding a skateboard today. He threw several newspapers in the front of several houses. However, he hit a pipe on the ground and sled on his feet. He was unable to get up due to his injuries.

Original Story Ending

His friend in the neighbor helped him get up and checked his wounds.

Alternative Ending 1 Reminder: ONE sentence; No significant overlap with the original ending; possible ending given the passage

The delivery man was riding a skateboard today. He threw several newspapers in the front of several houses. However, he hit a pipe on the ground and sled on his feet. He was unable to get up due to his injuries.

input goes here (ONE sentence)

Alternative Ending 2 Reminder: ONE sentence; No significant overlap with other endings; possible ending given the passage

The delivery man was riding a skateboard today. He threw several newspapers in the front of several houses. However, he hit a pipe on the ground and sled on his feet. He was unable to get up due to his injuries.

input goes here (ONE sentence)

Write alternative story endings!

Submit the endings

Figure 10: Interface used in the story ending writing task.

Instructions

We are collecting questions to study machine understanding of English. Thank you for your help! We appreciate your amazing jobs so far!

In this task, you will be given a passage and four options that have been collected as endings before and then selected automatically.

Your task is to **write two challenging questions with a single correct answer for each, given the short passage and the four answer options**. Accordingly, you need to **check the correct answer option for the questions you write** by checking the option, and the correct options to the two questions should be different from each other.

In writing the questions,

- [Important!] You must ensure that the answer to the question is **a single option among the four options presented**, and you **cannot modify the options**.
- The questions you write must be answerable based on the text in the passage and question, combined with some reasonings.
- You should assume that the reader has general background knowledge.
- You should also assume that the reader can speculate about something that is not explicitly stated in the text.
- It should not be easy to eliminate or guess the answer choice/s without reading the passage.
- Likewise, it should also not be easy to eliminate answer choice/s *without reading the question*.
- However, there is no problem in writing two similar questions (with minimal editions) as long as they have different options as their answers!

Guidelines for Writing Questions

Here are some guidelines for writing questions:

- Consider questions that ask about:
 - Characters' feelings and motivations (sad, happy etc consequence for the protagonists)
 - Causes and consequences of described events (both before and after the option occurs)
 - Definition, properties, and process explained in a passage (additional premises)
 - Summary and lesson of a passage
- Consider the following writing styles:
 - Use words and phrases that don't appear in the passage and the answer options
 - Use abstract words
 - Make construction of the sentence complex with negation or insertion
- Write questions that are answerable based only on the text in the passage and the question.

Figure 11: Instructions (1/3) used in the question writing task.

Examples

In the morning I got out of bed a little late. To make up for lost time I did not make a lunch. By the time I left the house I was only 2 minutes late. The traffic was light so I made it to work on time.

Good Example 1

Question: Which one of the following is most likely to happen after this, if I usually eat something before starting work?

Answer options:

- (a) I began waking up late often. (**correct**)
- (b) I had enough time to get a cup of coffee in the breakroom before starting work.
- (c) When I got to work I realized it was Saturday and I didn't need to be there.
- (d) I was even able to grab breakfast on the way to the office.

Explanation: the question has only one right answer. Using "after this" requires reading the passage. (When there are similar options, it happens that the possible question is such that one can assume the right answer without reading the passage. In that case, do your best to use paraphrasing to make it less easy. e.g. drink a cup of coffee --> need caffeine)

Good Example 2

Question: Which describes a habit I acquired from this?

Answer options:

- (a) I began waking up late often. (**correct**)
- (b) I had enough time to get a cup of coffee in the breakroom before starting work.
- (c) When I got to work I realized it was Saturday and I didn't need to be there.
- (d) I was even able to grab breakfast on the way to the office.

Explanation: the question has only one right answer. The word "habit" appears nowhere in either passage or answer options.

Other Good Question Examples

- Which one of the following most accurately expresses the conclusion drawn in the passage?
- Which one of the following most likely to happen after this if [insert condition]?
- Which one of the following most unlikely to occur after this when [insert condition]?
- What would be the most surprising/romantic/ironic ... outcome? (when it has only one ending applicable)
- What can we learn from this passage?
- What is the main theme of this passage?

Figure 12: Instructions (2/3) used in the question writing task.

Bad Example 1

Question: Which would be the most plausible story ending when I was not only on time but actually a couple of minutes early?

Answer options:

- (a) I began waking up late often.
- (b) I had enough time to get a cup of coffee in the breakroom before starting work. **(correct 1)**
- (c) When I got to work I realized it was Saturday and I didn't need to be there.
- (d) I was even able to grab breakfast on the way to the office. **(correct 2)**

Explanation: This question has multiple correct options.

Bad Example 2

Question: Which is the possible ending of the passage?

Answer options:

- (a) I began waking up late often. **(correct)**
- (b) I had enough time to get a cup of coffee in the breakroom before starting work. **(correct)**
- (c) When I got to work I realized it was Saturday and I didn't need to be there. **(correct)**
- (d) I was even able to grab breakfast on the way to the office. **(correct)**

Explanation: all the options are possible as a story ending.

Please refer to this FAQ for additional information.

Figure 13: Instructions (3/3) used in the question writing task.

Thanks for doing our HITs! We appreciate your excellent jobs so far. Given a passage and four answer choices, write two multiple-choice questions. Please make sure that for each question **there is only one correct answer** and avoid copying text directly from the passage and choices presented. Please read the full instructions before starting.

[Clarification 1] The answer options are expected to be the continuations to the passage, so yes/no questions usually do not fit well. We recommend creating questions with which or what.

[Clarification 2] You can split the question into two sentences, such as an additional condition + question. Again, we are infinitely grateful for what you have done with this challenging task!

▼ Passage 1 / 1

Megan went to the mall and saw a beautiful dress in a shopping window. She wanted the dress badly, but could not afford it. Megan decided to get a job and save up for the dress. After lots of hard work, she eventually saved up enough money.

Question 1

input goes here

Options (Please check one correct answer for Question 1)

- 1. Megan finally bought the dress and learned the value of hard work.
- 2. Upon leaving the store proud of her hard work paying off, Megan ran into a man who apologized for being clumsy, they pair ended up dating and later getting married.
- 3. Once she had saved up so much money, she decided to buy a nice ring with it instead.
- 4. She went back to the mall to get the dress but instead she found a different outfit that she loved even more.

Figure 14: Interface (1/2) used in the question writing task.

Question 2 (the correct answer should be different from the option already chosen for Question 1)

input goes here

Options (Please check one correct answer for Question 2)

1. Megan finally bought the dress and learned the value of hard work.

2. Upon leaving the store proud of her hard work paying off, Megan ran into a man who apologized for being clumsy, they pair ended up dating and later getting married.

3. Once she had saved up so much money, she decided to buy a nice ring with it instead.

4. She went back to the mall to get the dress but instead she found a different outfit that she loved even more.

If you think it is too difficult to come up with questions having one option as a single right answer, please provide us feedback! (e.g., some of the options are too similar)

Any of your insights will help us!

Get the completion code

Figure 15: Interface (2/2) used in the question writing task.

Instructions

We are collecting answers to the multiple choice questions to study machine understanding of English. Thank you for your help!

You will be presented with a passage and three sets of a question and options. The set of options is identical for the three questions.

1. **Select the one answer choice** given each question and a set of options. If you find the questions fall in the following conditions;

- Multiple choices are likely.
- No options are likely.
- Question is difficult to understand, comprehend, or answer. (Please rewrite the question in the box if possible)

choose the corresponding option among the three options indicating each condition. When nothing above represents the reason well, but you still find some problems in the presented questions, please indicate using the text box.

2. **Content Filtering:** If you notice any offensive words, negative stereotypes, or harmful social implications about social groups (such as people of color, women, immigrants, religions) in the presented passage, assumption, and story endings, please write them down in the textbox. You may also use this box to indicate any typos or grammatical errors. After we check if it is relevant, you will receive a bonus of \$0.3. There is no right answer to this, and as long as you don't write something completely irrelevant, you don't need to worry about if they are valid enough.

We recommend spending 3-5 minutes on each HIT.

Figure 16: Instructions used in the question validation task.