TellMeWhy: A Dataset for Answering Why-Questions in Narratives

Yash Kumar Lal
Stony Brook University
ylal@cs.stonybrook.edu

Nathanael Chambers
US Naval Academy
nchamber@usna.edu

Raymond Mooney
University of Texas, Austin
mooney@cs.utexas.edu

Niranjan Balasubramanian
Stony Brook University
niranjan@cs.stonybrook.edu

Abstract

Answering questions about why characters perform certain actions is central to understanding and reasoning about narratives. Despite recent progress in QA, it is not clear if existing models have the ability to answer "why" questions that may require commonsense knowledge external to the input narrative. In this work, we introduce **TellMeWhy**, a new crowd-sourced dataset that consists of more than 30k questions and free-form answers concerning why characters in short narratives perform the actions described. For a third of this dataset, the answers are not present within the narrative. Given the limitations of automated evaluation for this task, we also present a systematized human evaluation interface for this dataset. Our evaluation of state-of-the-art models show that they are far below human performance on answering such questions. They are especially worse on questions whose answers are external to the narrative, thus providing a challenge for future QA and narrative understanding research.

1 Introduction

The actions people perform are steps of plans to achieve their desired goals. When interpreting language, humans naturally understand the reasons behind described actions, even when the reasons are left unstated (Schank and Abelson, 1975). For NLP systems, answering questions about *why* people perform actions in a narrative can test this ability. Answering such questions often requires filling the implicit gaps in the story itself.

Consider this narrative from ROCStories (Mostafazadeh et al., 2016b):

Rudy was convinced that bottled waters all tasted the same. He went to the store and bought several popular brands. He went back home and set them all on a table. He spent several hours tasting them one by one. He came to the conclusion that they actually did taste different.

Now try to answer the question, *"Why did he go to the store and buy several popular brands?"* The answer *"he wanted to taste test"* is not explicit in the narrative and requires us to read between the lines to fill in the gaps (Norvig, 1987). While humans can visualise and process the events in a story to hypothesize why they might have occurred (Kintsch and Dijk, 1978), current NLP systems fall well short of exhibiting similar capabilities. They are unable to adequately formulate the reasons behind actions in specific contexts.

How can we get NLP models to reason about why actions are performed? One way is to consider theories like script learning (Schank and Abelson, 1975; Pichotta and Mooney, 2014) or learning from co-occurrence (Chambers and Jurafsky, 2009). But they only partially capture this type of knowledge – much like other forms of commonsense knowledge, the reasons for why actions are performed are often left implicit in text. Even though there are many large scale QA datasets, they rarely contain questions about *why* people perform actions.

Therefore, we introduce the TellMeWhy dataset, a collection of 30,519 such why-questions, each with 3 "gold standard" human answers. Each record in TellMeWhy contains a short story, an associated question, and its 3 possible answers.

Further, we focus on enabling *human* evaluation of this dataset; human evaluation is more reliable than automatic metrics to evaluate such systems (Celikyilmaz et al., 2020; Gatt and Krahmer, 2018). However, reliability of human judgment is substantially impacted by experimental setup (Novikova et al., 2018; Santhanam and Shaikh, 2019). There is little consensus on how human evaluations should be conducted, so results are often incomparable across evaluations.

To this end, we present a systematized evaluation framework on MTurk for the TellMeWhy text generation task – and release the framework for future

researchers. The MTurk interface asks annotators to rate generated answers on their grammaticality and validity. We show that with our interface human answers are judged to be of high quality (99% grammatical, 96% valid) with strong inter-annotator agreement at 0.88 Fleiss Kappa. This indicates high agreement and also confirms the design of our interface.

Finally, we present baseline results for TellMeWhy and compare against our human ceiling. We finetune two large language models that have proven to be effective for a variety of tasks, GPT-2 (Radford et al., 2019) and T5 (Raffel et al., 2020), and a dedicated question answering model, UnifiedQA (Khashabi et al., 2020), to perform this task. Human evaluation is performed on their outputs from independent test data. All models significantly under-perform the human benchmark and are especially worse on questions where the answer cannot be simply copied over from text in the narrative. The results clearly demonstrate the difficulty for current models to convincingly answer such why-questions.

This paper's contributions are as follows: (1) we introduce TellMeWhy, a large dataset of English why-questions for narratives derived from ROCStories (Mostafazadeh et al., 2016a) and CATERS (Mostafazadeh et al., 2016b) along with answers from 3 distinct humans, (2) a systematized human evaluation interface to calibrate model outputs consistently, and (3) show that current models are ill-equipped to perform this task. We release the dataset and human evaluation suite at `http://lunr.cs.stonybrook.edu/tellmewhy`.

2 Related Work

2.1 Datasets containing why-questions

Most of the datasets related to why-questions fall into one or more of the following categories: (1) very small size, (2) not focused on stories, or (3) focused on connecting known events instead of answering reasoning questions.

Some corpora of why-questions have been collected manually: corpora described in Verberne et al. (2006) and Verberne et al. (2007) both comprise fewer than 400 questions and corresponding answers (one or two per question) formulated by native speakers. Dunietz et al. (2020) demonstrate that it is important to define what we want models to comprehend when building datasets for machine reading comprehension (MRC) tasks. They design templates of understanding corresponding to the four elements identified by Zwaan et al. (1995). For 201 questions, they design multiple-choice questions derived from (Lai et al., 2017) to test understanding of different categories of events. All of these are very small corpora that cannot be viably used to further a model's understanding of why-questions in stories.

Higashinaka and Isozaki (2008) extend an existing factoid QA system to answer why-questions by integrating corpus based features, calling it NAZEQA. Oh et al. (2012) extract a set of answer candidates from a web corpus, and perform re-ranking using SVMs to predict the right answer. Oh et al. (2019) use an adversarial learning framework to generate a vector representation from the passage to judge whether the passage actually answers the why-question. These papers focus on Japanese news (Fukumoto et al., 2007; Oh et al., 2012), including NTCIR-6, and most critically, all these datasets are very small.

Some prior work focuses on knowledge extraction, not the *reasons* behind the actions. Mrozinski et al. (2008) built a corpus of why-questions related to Wikipedia articles. These were general knowledge questions with solicited answers from paid workers. Dependency parsing can be used to rephrase why-questions into statements with a 'because' prompt to elicit explanations from models (Nie et al., 2019). PhotoshopQuiA (Dulceanu et al., 2018) contains questions and answers specifically about Photoshop.

NarrativeQA (Kočiský et al., 2018) provides a dataset of 1,567 stories (books and movie scripts) containing 46,765 wh-questions written and answered by human annotators. Unfortunately, only 9.78% are why-questions, which makes for a small collection. QuAIL (Rogers et al., 2020) has a small subset of multiple choice questions pertaining to causality in user stories. These datasets are targeted at broad abilities of reading comprehension, not specifically about explaining actions in stories.

Some recent datasets causally connect events in text, but they do not target answering why-questions. ATOMIC (Sap et al., 2019) consists of entries that describe a likely cause/effect of events. Most notably, ATOMIC is *non-contextual* so it is more about general knowledge, not interpreting a specific story/context. Perhaps most relevant is GLUCOSE (Mostafazadeh et al., 2020), a crowdsourced dataset of implicit commonsense

Dataset	Size	Domain
NTCIR-6	200	Japanese news
Mrozinski et al. (2008)	695	Wikipedia
PhotoshopQuiA	2,854	Product focused
NarrativeQA	4,573	Books+Movie scripts
Dunietz et al. (2020)	201	Exam questions

Table 1: Previous why-question corpora. NarrativeQA has 46,765 questions of which 4,573 are why-questions.

knowledge in the form of causal mini-theories grounded in narrative context. These theories are semi-structured inference rules. This dataset is not aimed at answering why-questions, but at creating direct relationships between events already mentioned in the story. They focus on capturing specific cause-enable type relations. Annotators were given a very constrained task – they had to select options from a drop down menu describing inference rules.

Abductive commonsense reasoning tests whether models can come up with a plausible explanation to connect a set of events. Bhagavatula et al. (2020) present ART with two abductive tasks: 1) given two observations, select one out of two plausible hypotheses, 2) and generate text connecting two events. This line of work focuses on connecting the dots between two events and does not address explaining *why* an action was performed. Our work crucially differs from these because the answer is often not in the story at all. StrategyQA (Geva et al., 2021) is a new dataset focusing on performing better implicit reasoning for multi-hop question answering tasks.

We summarize the different why-questions corpora in Table 1. None of them represent a large dataset focused on answering why-questions about actions in a narrative.

2.2 Human evaluation for NLG tasks

Among language generation tasks, machine translation has received the most attention in terms of human evaluation. Qualified crowd workers score output translations given the source or reference text to calibrate MT systems (Sakaguchi and Van Durme, 2018; Graham et al., 2013, 2014). WMT conducts annual evaluation of outputs of systems submitted to the shared task and uses it as one of the primary metrics (along with BLEU) to rank systems (Bojar et al., 2016, 2017, 2018; Barrault et al., 2019, 2020).

ChatEval (Sedoc et al., 2019) is an evaluation platform for chatbots. Zellers et al. (2020) present a leaderboard for their advice generation task. These platforms incorporate some manual analysis, but focus on very different tasks. None of their Mechanical Turk interfaces can be used for our task. We were unable to find a consistent interface for human evaluation of an open-ended question answering task. To address this flaw, we propose a standard human intelligence task (HIT) evaluation scheme for our dataset.

3 Dataset Creation

We want to test the abilities of models to understand the reasoning behind actions in a story. Therefore, we create a dataset of *why* questions that ask for explanations for actions performed in a story. Answering these questions requires an understanding of the events that are explicit in the story as well as access to implicit common-sense knowledge on how people use actions as parts of plans to achieve goals. To cover a wide-range of common situations, we utilize ROCStories (Mostafazadeh et al., 2016a), a collection of 45,496 five-sentence commonsense stories. We also develop a small "hidden" test set that was only used for the final evaluation using the CATERS (Mostafazadeh et al., 2016b) subset of ROCStories.

3.1 Why-Question Generation

Our strategy for creating *why* questions is simple. For each action in the narrative, we formulate a why question by applying simple template-based transformations. We dependency parse each sentence using SpaCy's en_core_web_sm model (Honnibal et al., 2020). We use the generated parse tree to rephrase the sentence into a question about the action described. The generated parse tree is used to extract the subject, object, and verb. We consider 3 types of sentences and design question templates accordingly: (1) sentences that have a primary and auxiliary verb, (2) sentences that only have a primary verb, and (3) sentences that only contain an auxiliary verb. For the first, the question template is: "Why {aux_verb} {subject} {verb} {object}?", for the second: "Why did {subject} {verb_lemma} {obj}?", and for the third; "Why {aux_verb} {subject} {obj}?".

This procedure yielded a little over 113k questions from ROCStories, and 489 questions from the CATERS portion. We selected at random 32,165 questions from stories that had at least three ques-

Split	# stories	# questions
Train	7558	23964
Dev	944	2992
Test	944	3099
Hidden Test	190	464
Total	9,636	30,519

Table 2: Dataset Statistics

tions[1]. We ensure that there is no overlap between the two subsets.

3.2 Collecting Answers

We crowd-sourced answers to these questions using Amazon Mechanical Turk. Figure 1a shows the interface used to collect these answers. Annotators were presented a narrative and asked to answer three *why* questions in free-form. For each question, they were also asked to provide judgments about the comprehensibility of the question, and whether the narrative explicitly contained the answer. They also selected the sentences from the narrative which influenced their answer (if any). To avoid variability in answer prefixes, we provide a prompt to start answering the question. We rephrase the sentence from which the question was generated to create these prompts. We consider the same categorisation of sentences described in subsection 3.1. For sentences that have both primary and auxiliary verbs, the answer prompt is of the form: "{subject} {aux_verb} {verb} {object} because...". When it only contains a primary verb, it is of the form: "{subject} {verb} {object} because...". If it only contains an auxiliary verb, it is of the form: "{subject} {aux_verb} {object} because...". We found, over several iterations of this HIT, that providing a prompt gave workers an initial direction and improved the quality of answers collected.

We ask three distinct annotators (three-way redundant task) to answer each of these questions. Annotators are not allowed to copy pieces of text to make up an answer. We discard questions that were deemed incomprehensible by any annotator.[2] With this process, we obtained 3 answers each from 30,055 questions from 9,636 stories (see Appendix B for more details). Table 2 shows basic statistics of the dataset. We refer to annotations from the CATERS data as the hidden test set. Examples of records in the dataset are presented in Table 3. The narrative does not explicitly contain an answer for the second question. We call these types the implicit-answer questions; they require extra common-sense inference to produce a plausible answer. Questions are categorised as implicit-answer if at least 2 out of 3 human annotators indicate that the answer cannot be explicitly found in the narrative. The annotators indicated as much and, based on their commonsense knowledge, provided plausible answers.

Story: Sandra got a job at the zoo. She loved coming to work and seeing all of the animals. Sandra went to look at the polar bears during her lunch break. She watched them eat fish and jump in and out of the water. She took pictures and shared them with her friends. **Question:** Why did Sandra go to look at the polar bears during her lunch break? **Ans:** she wanted to take some pictures of them.
Story: Cam ordered a pizza and took it home. He opened the box to take out a slice. Cam discovered that the store did not cut the pizza for him. He looked for his pizza cutter but did not find it. He had to use his chef knife to cut a slice. **Question:** Why did Cam order a pizza? **Ans:** Cam was hungry.

Table 3: TellMeWhy examples. The first is answerable directly from text in the story, but the second requires external knowledge. We only show one out of three available answers here.

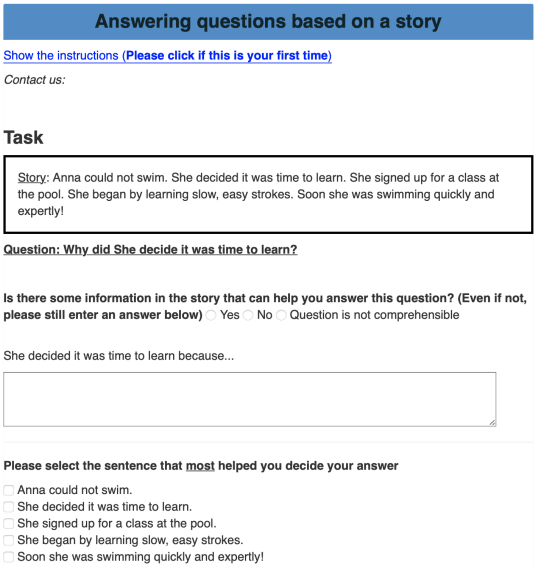
3.3 Validating Answers

To ensure an even higher-quality test set, we conducted another round of crowdsourcing to validate the answers by the first set of crowd-workers on the CATERS portion (464 questions). This validation interface is show in Figure 1b. It also serves as the base design for our systematized human evaluation. Annotators are presented a story, a related question, and the three answers that were collected as described in Section 3.2.

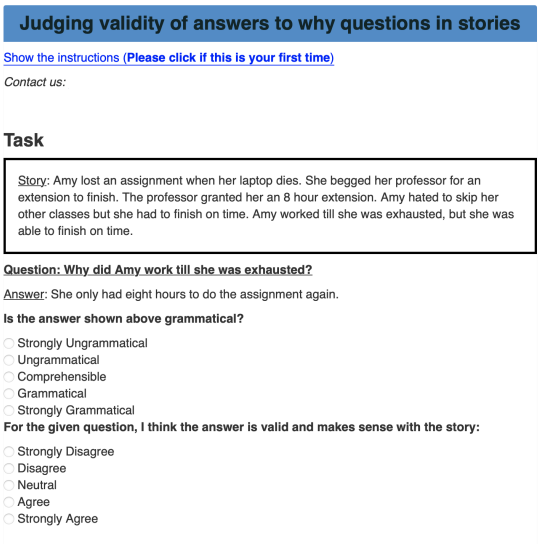
Three new annotators then rated two aspects of each answer:
(1) **Grammaticality** – Workers are asked to rate

[1]Since we ask annotators to read an entire story to answer these questions, avoiding stories with fewer questions optimizes reading time.

[2]On ROCStories we discarded 1,546 questions and on CATERS we discarded 25 questions



(a) Task 1: Answer collection



(b) Task 2: Answer validation

Figure 1: MTurk interfaces used to curate data from crowd-source workers

the grammaticality of each answer on 5-point Likert scales, ranging from 'Strongly Ungrammatical' to 'Strongly Grammatical'. An answer is strongly grammatical if it follows all the rules of English grammar. It is grammatical if there is a mistake in tense, number, punctuation or something minor. It is comprehensible if there are clear grammatical mistakes but its meaning can be inferred, and it is then considered to be neutral on the Likert scale.

(2) **Validity** – Workers are asked to rate the validity of each answer on a 5-point Likert scale. Given the story and question, the annotators check if the given answer 'is valid and makes sense with the story'. An answer is considered invalid if it does not give a plausible reason relevant to the question asked and instead states irrelevant information.

Annotators agreed (by majority) that 99.07% of answers are grammatical and 95.47% of answers are valid. On grammaticality, there is some disagreement in judgment 0.7% of the time, while there is some disagreement in judgment 1% of the time for answer validity. We measured the inter-rater reliability of annotators' judgments using weighted Fleiss's Kappa (Marasini et al., 2016) and follow the weighting scheme used by Bastan et al. (2020). This measure has a penalty for each dissimilar classification based on the distance between two classes. For instance, if two annotators classify a document as a positive, the agreement weight is 1, but if one classifies as a positive, and the other classifies as slightly positive the agreement weight is less. The weighted agreement score for this subset is 0.88 for grammaticality annotations and 0.81 for validity annotations, indicating that the annotations are highly reliable. More details can be found in Appendix C.2.

4 Dataset Analysis

One of the key distinguishing aspects of answering *why* questions is that, in addition to understanding explicitly stated events, they also require access to commonsense explanations that may be external to the narrative. We conduct some analyses to investigate the prevalence of this phenomenon: (i) We asked annotators to judge whether the answer to a question could be found stated explicitly or only implicitly in the narrative and find that at least two out of three annotators could not find explicit answers in the story 28.82% of the time. (ii) We also asked crowd-workers to indicate which sentences helped them answer the question. Out of 91,557 collected answers, we find that 39,661 answers were provided without an influential sentence from the story. (iii) Last, we observe that there is only a 57.04% lexical overlap between the words used in answers and the original narrative. This suggests that annotators included new inferred information in their answers, instead of just copying something from the story. We calculate lexical overlap as the number of common tokens in the narrative and the answer divided by the length of the answer.

We hypothesize that questions about the first

action in a story are more difficult to answer since there is no prior information to provide an explicit answer. We find that 55.03% of such questions were judged to be implicit-answer questions by a majority of the assigned annotators. Such questions help test systems' ability to infer plausible answers rather than just copy answers from the text.

We also evaluated the diversity of the answers for each question using simple lexical overlap. Of the 30k questions, only 150 questions had over 90% overlap in all 3 answers, i.e., essentially, the 3 distinct annotators wrote the same answer. For 4,243 other questions, two out of three answers had over 90% overlap. But for the vast majority of 26,068 questions, we obtained 3 fairly diverse answers. The average overlap between them is 26.12%. On average, the answers were 7.59 words long.

Overall, this analysis indicates how TellMeWhy differs from prior datasets. The answers cannot always be retrieved or connected to other events in the given text.

5 Benchmarking

How well do large language models answer *why* questions on narratives and what are their failure modes? To answer these, we use TellMeWhy to benchmark the performance of multiple state-of-the-art models and provide an analysis of their performance.

Formally, given a story S as context and a related why-question Q, models are required to generate a plausible answer A for the question. Since the answers are open-ended texts we compare them on standard automatic evaluation metrics for generation but also conduct a human evaluation.

5.1 QA Models

GPT-2 (Radford et al., 2019) is a large transformer-based language model trained on an enormous web corpus, which has been shown to be effective on a wide-range of language related tasks including question answering. It was one of the first models trained on diverse data to outperform domain-specific language models.

We used Huggingface (Wolf et al., 2020) to finetune a pretrained GPT-2 model on TellMeWhy. As input, the model receives a concatenation of the narrative and the related question (in that order), and the target is the answer. The input and target are separated using the '[SEP]' token. We finetune the model with batch size 16, learning rate 1e-5 and maximum output length 25. The model is trained until the dev loss fails to improve for 3 iterations.

T5 (Raffel et al., 2020) is an encoder-decoder model pre-trained on a mixture of unsupervised and supervised tasks in a multi-task setting, where each task is converted into a text-to-text format. It is a text-to-text model, which means it can be trained on arbitrary tasks involving textual input and output. T5 has achieved the state of the art on many natural language understanding (NLU) tasks. More details about hyperparameter sweeps can be found in Appendix A.

We finetuned a pretrained T5-base model from HuggingFace (Wolf et al., 2020) on TellMeWhy. Since it is a natural language generation task related to a story, we use the SQuAD format specified in Appendix D.15 of Raffel et al. (2020) to format our inputs. Our narrative serves as the 'context' and the why-question is used as the 'question' in the selected input format. We train the model with batch size 16, learning rate 5e-5, maximum source length 75 and maximum answer length 30. The model is trained until the dev loss fails to improve for 3 iterations.

UnifiedQA (Khashabi et al., 2020) is a single pre-trained model that performs well across 20 different question answering datasets. It is built on top of a T5 model and simplifies finetuning by unifying the various formats used by T5. Its ability to perform both extractive and abstractive QA tasks makes it a suitable candidate for calibrating this task. A pretrained version of this model is available via HuggingFace (Wolf et al., 2020) under the name "allenai/unifiedqa-t5-base". The input format for this model is simple, just requiring the question and the narrative to be separated by a newline symbol. We train this model using learning rate 1e-5 (same as the original paper) and retain other hyperparameters from finetuning T5 as described above.

5.2 Automatic Evaluation

We evaluate all of the above models on both the test set and the hidden test set (questions from CATERS data). For automatic evaluation, we report BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), BLEURT (Sellam et al., 2020) scores using the bluert-base-128 checkpoint, and BertScore (Zhang* et al., 2020) using the default roberta-large checkpoint. These numbers are presented in Table 4.

Evaluated on	Model	BLEU	RG-L F1	BLEURT	BertScore
Full Test Set	GPT-2-OO	0.04	0.07	-1.30	0.08
	GPT2-FT	1.3	0.13	-0.96	0.11
	T5-OO	5.67	0.13	-1.20	0.14
	T5-FT	**13.33**	0.24	-0.70	**0.34**
	UnifiedQA	13.03	**0.25**	**-0.71**	0.30
Implicit-Answer Qs in Test Set	GPT-2-OO	0.07	0.06	-1.30	0.05
	GPT2-FT	1.39	0.12	-1.02	0.09
	T5-OO	3.12	0.11	-1.24	0.12
	T5-FT	**7.27**	0.17	**-0.89**	**0.27**
	UnifiedQA	6.63	**0.18**	-0.89	0.24

Table 4: Performance of models on the full test set and on implicit-answer questions in the test set using automated metrics. RG-L denotes ROUGE-L. The OO suffix denotes the vanilla version of the model while the FT version denotes the finetuned version.

We select one human answer at a time and (using SacreBLEU (Post, 2018)) calculate the BLEU scores for model output with all three references, and select the maximum. Since BLEURT is a sentence level metric, to calculate the reported BLEURT, we average all the (output, reference) scores to obtain a corpus score for each reference. We then select the maximum BLEURT corpus score over all 3 human references. It is important to note that BLEURT was proposed as a metric for relative comparison, not absolute calibration. We also report BertScore F1[3] (Zhang* et al., 2020) as another semantic automatic evaluation metric. We report a max BertScore in the same way as BLEURT and BLEU: by taking the maximum score of the model output with each human answer taken one at a time.

Vanilla model results are obtained by loading an existing pretrained model from HuggingFace and running inference with the input formats described above. They are not trained on TellMeWhy. We see that vanilla pretrained models are unable to perform this task at all. Finetuning a pretrained model results in improvements since it better models the relationship between the story, the question, and a possible answer. On the full test set, both the finetuned T5 and the UnifiedQA model perform the best on our task. However, the overall performance of these models remains poor. In Table 4, we also see that models perform a lot worse on implicit-answer questions.

[3]idf and rescale_with_baseline flags are set to True.

5.3 Human Evaluation

For open-ended text generation tasks like answering why-questions, the absence of an automatic evaluation that correlates well with human judgments is a major challenge (Chen et al., 2019; Ma et al., 2019; Caglayan et al., 2020; Howcroft et al., 2020).

We conduct a human evaluation on the hidden test set with a standardized interface to compare different models. We want to measure whether a model produces coherent and grammatical output and more importantly, whether the produced output is a valid answer for the given question. Our validation HIT subsection 3.3 showed a way to conduct human evaluation of answers provided by other crowd-workers. We modified this HIT design to evaluate generated answers from models. For a given question, we present just one answer from a single model and then ask the crowd-workers to assess its grammaticality and validity.

For each story, question, and a model's answer, we ask 3 distinct annotators to provide judgments about grammaticality and validity. This serves as the human evaluation interface for our task. A sample HIT can be seen in Figure 1b.

We perform human evaluation of the fine-tuned versions of T5 and UnifiedQA, the two models that performed the best on automatic metrics. We evaluate the outputs of these models on the hidden test set. We calculated inter-annotator agreement for these judgments using the method described in subsection 3.3, and they were >80%, indicating high agreement. The models mostly produce grammatical answers, but fail to adequately ex-

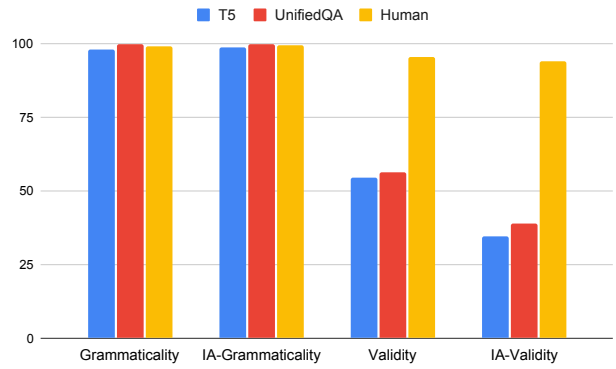


Figure 2: Human evaluated performance of answers. The IA prefix denotes performance on implicit-answer questions in the data.

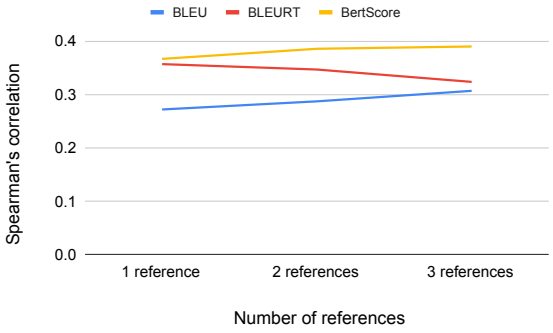


Figure 3: Correlation of automatic metrics with human validity judgment of model outputs. For each question, we have 3 crowdsourced human answers available to us. We selected a number of human answers randomly (X-axis) and calculated scores for each model output across different automatic metrics. Finally, we obtained Spearman's correlation (Y-axis) of these scores in comparison with Likert judgments provided by annotators for each human answer.

plain many actions in the story. Figure 2 shows that, under human evaluation, models significantly under-perform humans at producing valid answers to why-questions. Models fare worse when the answers to the questions are external to the narrative.

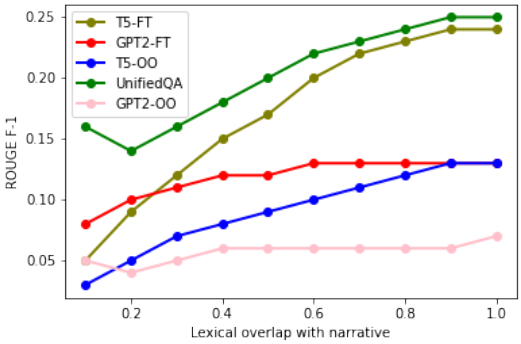
Human evaluation is slow and expensive, so we performed a correlation analysis between the automated metrics and human judgments to gauge usefulness of popular automated metrics. Figure 3 shows that the embedding-based metrics are only weakly correlated with human validity judgments, while lexical metrics did even worse. None of the automatic metrics show a strong correlation, confirming our earlier assertion that human evaluation is the most appropriate way to analyze model performance on this open-ended generation task. BertScore has at least a moderate correlation with human validity judgments, and is therefore arguably the most useful for rapidly evaluating models during development. BLEU and BertScore improve their correlation with human judgments slightly as the number of human reference answers is increased; however, the increase is somewhat disappointing. Inexplicably, BLEURT's correlation actually decreases slightly with increasing human references, raising additional questions with respect to utilizing this metric. One possibility is that, by using an increasing number of references and taking the maximum score, BLEURT might overestimate the quality of answers as compared to human judgments.

Our human judgment interface can serve as a standard human evaluation of any future model's performance on our dataset, and we will make code available for automatically generating HITs for evaluating the outputs of any model. This standardized evaluation approach is similar in spirit to GENIE (Khashabi et al., 2021), a contemporary work that also presents an evaluation framework for a large set of generation tasks.

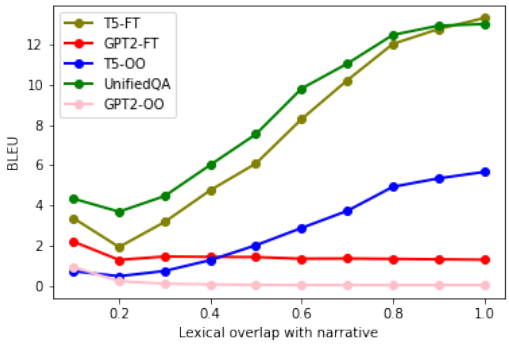
5.4 Analysis

In order to better understand when models are generating valid answers, we analyzed the correlation between model performance and a proxy for checking when human provided answers were in the input narrative. To this end, we aligned ROUGE F-1 scores with the lexical overlap of human answers and the story text. Figure 4a shows how ROUGE F-1 scores for our models increases as the lexical overlap also increases between the answers and corresponding story. The same is presented for BLEU in Figure 4b. Perhaps not surprisingly, this empirically shows that models do best when the answer is in the text, and suffer greatly when it is not (implicit answers). This further illustrates the value of TellMeWhy, as well as its challenge, that standard models are largely incapable of performing the reasoning needed to produce plausible answers that are assumed common knowledge by the story writer.

Table 5 also shows that the best performing models mainly learnt to copy complete or parts from the narrative to generate answers, treating this largely as an extractive task. On average, more than three-fourths of T5 and UnifiedQA's answers are



(a) ROUGE F-1 trend



(b) BLEU trend

Figure 4: Model performance on different metrics with change in lexical overlap between a question's answers (as provided by humans) and the related narrative.

System	Copied ans	Avg overlap
Vanilla GPT2	0%	23.09%
Finetuned GPT2	0%	28.94%
Vanilla T5	5.50%	53.71%
Finetuned T5	59.44%	85.91%
UnifiedQA	27.44%	76.51%
Human Answers	35.03%	57.04%

Table 5: Overlap between answers and the original narrative. This indicate how much original text models produce.

based on words in the narrative text. T5 is worse compared to UnifiedQA in terms of copying, with a much larger fraction of questions (59.44% vs 27.44) with high lexical overlap (i.e. lexical overlap > 90%). In comparison, the average narrative overlap for human answers is much lower than the best-performing models, since people are able to infer answers that are not in the text. If the models are to successfully answer *why* questions, they need to look beyond copying texts.

6 Conclusion

This paper introduces a large, novel QA dataset, **TellMeWhy**, containing questions about *why* characters in a narrative perform their depicted actions. This challenge problem complements the variety of existing QA datasets, addressing the scarcity of "why" questions. Using both automated metrics and human evaluation, we show that existing deep-learned language models perform quite poorly at answering such questions. We also illustrate the uniqueness of this challenge where the answer is sometimes in the story itself, but often not, thus requiring a richer model that can draw on commonsense knowledge or external reasoning abilities.

We believe that progress on answering such questions requires new systems that can reason about actions, plans, and goals in order to achieve a deeper understanding of narrative text, as was initially argued over four decades ago (Schank and Abelson, 1977). We hope that TellMeWhy encourages further research in this area.

Acknowledgement

This material is based on research that is supported in part by the Air Force Research Laboratory (AFRL), DARPA, for the KAIROS program under agreement number FA8750-19-2-1003 and in part by the National Science Foundation under the award IIS #2007290. The authors would like to thank the anonymous reviewers and the area chair for their feedback on this work. We would also like to thank Horace Liu for helping us run some experiments for the camera ready version.

References

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019.

Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Mohaddeseh Bastan, Mahnaz Koupaee, Youngseo Son, Richard Sicoli, and Niranjan Balasubramanian. 2020. Author's sentiment prediction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 604–615, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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*.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Ozan Caglayan, Pranava Madhyastha, and Lucia Specia. 2020. Curious case of language generation evaluation metrics: A cautionary tale. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2322–2328, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 602–610, Suntec, Singapore. Association for Computational Linguistics.

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Evaluating question answering evaluation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 119–124, Hong Kong, China. Association for Computational Linguistics.

Andrei Dulceanu, Thang Le Dinh, Walter Chang, Trung Bui, Doo Soon Kim, Manh Chien Vu, and Seokhwan Kim. 2018. PhotoshopQuiA: A corpus of non-factoid questions and answers for why-question answering. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jesse Dunietz, Greg Burnham, Akash Bharadwaj, Owen Rambow, Jennifer Chu-Carroll, and Dave Ferrucci. 2020. To test machine comprehension, start by defining comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7839–7859, Online. Association for Computational Linguistics.

J. Fukumoto, T. Kato, F. Masui, and T. Mori. 2007. An overview of the 4th question answering challenge (qac-4) at ntcir workshop 6. In *NTCIR*.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *J. Artif. Int. Res.*, 61(1):65–170.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2014. Is machine translation getting better over time? In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 443–451, Gothenburg, Sweden. Association for Computational Linguistics.

Ryuichiro Higashinaka and Hideki Isozaki. 2008. Corpus-based question answering for why-questions. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. 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.

Daniel Khashabi, Gabriel Stanovsky, Jonathan Bragg, Nicholas Lourie, Jungo Kasai, Yejin Choi, Noah A. Smith, and Daniel S. Weld. 2021. Genie: A leaderboard for human-in-the-loop evaluation of text generation.

W. Kintsch and T. Dijk. 1978. Toward a model of text comprehension and production. *Psychological Review*, 85:363–394.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

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.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

D. Marasini, P. Quatto, and E. Ripamonti. 2016. Assessing the inter-rater agreement for ordinal data through weighted indexes. *Statistical Methods in Medical Research*, 25:2611 – 2633.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016a. 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, Alyson Grealish, Nathanael Chambers, James Allen, and Lucy Vanderwende. 2016b. CaTeRS: Causal and temporal relation scheme for semantic annotation of event structures. In *Proceedings of the Fourth Workshop on Events*, pages 51–61, 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.

Joanna Mrozinski, Edward Whittaker, and Sadaoki Furui. 2008. Collecting a why-question corpus for development and evaluation of an automatic QA-system. In *Proceedings of ACL-08: HLT*, pages 443–451, Columbus, Ohio. Association for Computational Linguistics.

Allen Nie, Erin Bennett, and Noah Goodman. 2019. Learning to explain: Answering why-questions via rephrasing. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 113–120, Florence, Italy. Association for Computational Linguistics.

Peter Norvig. 1987. Inference in text understanding. In *Proceedings of the Sixth National Conference on Artificial Intelligence - Volume 2*, AAAI'87, page 561–565. AAAI Press.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2018. RankME: Reliable human ratings for natural language generation. 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 72–78, New Orleans, Louisiana. Association for Computational Linguistics.

Jong-Hoon Oh, Kazuma Kadowaki, Julien Kloetzer, Ryu Iida, and Kentaro Torisawa. 2019. Open-domain why-question answering with adversarial learning to encode answer texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4227–4237, Florence, Italy. Association for Computational Linguistics.

Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Takuya Kawada, Stijn De Saeger, Jun'ichi Kazama, and Yiou Wang. 2012. Why question answering using sentiment analysis and word classes. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 368–378, Jeju Island, Korea. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Karl Pichotta and Raymond Mooney. 2014. Statistical script learning with multi-argument events. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 220–229, Gothenburg, Sweden. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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.

Keisuke Sakaguchi and Benjamin Van Durme. 2018. Efficient online scalar annotation with bounded support. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 208–218, Melbourne, Australia. Association for Computational Linguistics.

Sashank Santhanam and Samira Shaikh. 2019. Towards best experiment design for evaluating dialogue system output. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 88–94, Tokyo, Japan. Association for Computational Linguistics.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3027–3035.

R. Schank and R. P. Abelson. 1977. Scripts, plans, goals and understanding: an inquiry into human knowledge structures.

Roger C. Schank and Robert P. Abelson. 1975. Scripts, plans, and knowledge. In *Proceedings of the 4th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'75, page 151–157, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

João Sedoc, Daphne Ippolito, Arun Kirubarajan, Jai Thirani, Lyle Ungar, and Chris Callison-Burch. 2019. ChatEval: A tool for chatbot evaluation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 60–65, Minneapolis, Minnesota. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Suzan Verberne, Lou Boves, Nelleke Oostdijk, and Peter-Arno Coppen. 2006. Data for question answering: The case of why. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Suzan Verberne, Lou Boves, Nelleke Oostdijk, and Peter-Arno Coppen. 2007. Evaluating discourse-based answer extraction for why-question answering. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, page 735–736, New York, NY, USA. Association for Computing Machinery.

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, Elizabeth Clark, Lianhui Qin, Ali Farhadi, and Yejin Choi. 2020. Evaluating machines by their real-world language use. *arXiv preprint arXiv:2004.03607*.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Rolf A. Zwaan, Mark C. Langston, and Arthur C. Graesser. 1995. The construction of situation models in narrative comprehension: An event-indexing model. *Psychological Science*, 6(5):292–297.

A Hyperparameter Sweep

We describe the hyperparameters and the range of values we experimented with. The best hyperparameters are chosen on the basis of model loss on the validation set. For both GPT-2 (Radford et al., 2019) and T5 (Raffel et al., 2020), we conduct guided sweeps for learning rate, batch size and epochs. We experiment with 1e-5, 5e-5 and 1e-4 for learning rate. Batch sizes of 8, 16 and 32 were tried. Models were trained for 20, 30 and 50 epochs, and we found that models converged between 30 and 50 epochs. In the case of T5, we also experiment with different lengths of inputs and target outputs. We trained models with maximum source lengths of 50, 60 and 75 tokens. For target length, we experimented with 15, 25 and 30 tokens. The maximum output length is treated as a hyperparameter for GPT-2, and we tried 15, 20, 25 and 30 tokens.

B Dataset Creation

The method described in subsection 3.1 creates 489 questions from the 200 stories in the CATERS dataset – 36 stories with 1 question, 63 with 2, 59 with 3, 30 with 4, and 6 with 5. We collect 3 human answers for all questions. For ROCStories, this creates 113,213 questions from 45,496 stories – 7,555 stories with 1 question, 13,431 with 2, 13,349 with 3, 7356 with 4, and 1865 with 5. We randomly select 32,165 questions from stories with 3 or 5 questions, for ease and efficiency of collecting annotations. This is the smallest number for which we could gather 3 answers for at least 30,000 questions, which is a reasonable-sized dataset for training or fine-tuning large NLP models.

C Mechanical Turk tasks

C.1 Instructions

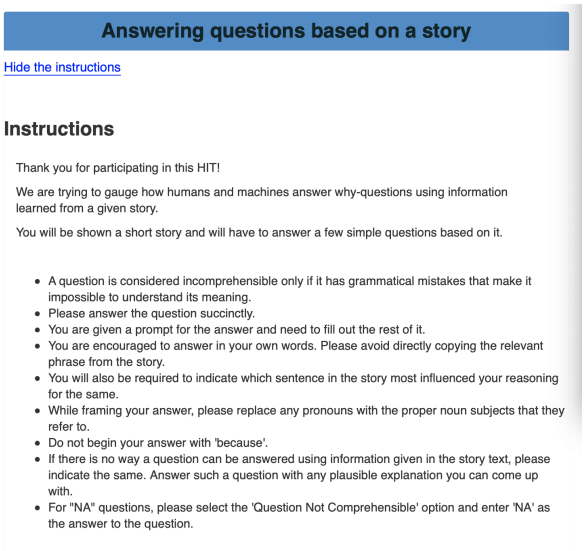
We present the instructions given to annotators for both the tasks in Figure 5. Annotators were given clear direction for both tasks. We restricted both tasks to master turkers. The second task (answer validity) was also used a sanity check for answers collected in the first task (answer collection). Using results of the answer validity task (mentioned in subsection 3.3), we see that humans provided high quality answers in the answer curation task.

C.2 Inter-annotator agreement

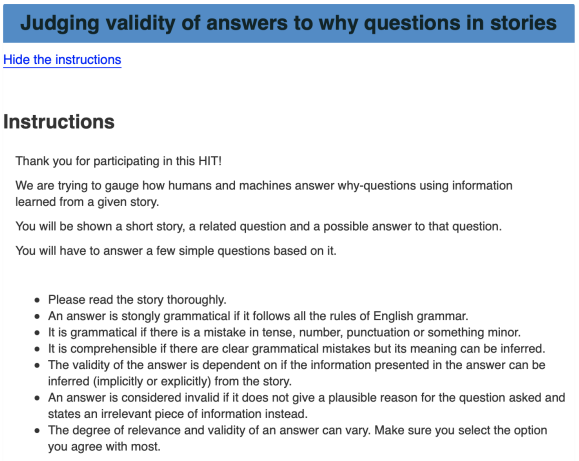
We use weighted Fleiss Kappa to calculate inter-rater reliability. The weights between different classes are shown in Table 6 where negative, slightly negative, neutral, slightly positive, and positive classes are shown with -2, -1, 0, 1, and 2. We follow the setup used in Bastan et al. (2020) for a similar multi-class labeling task.

	-2	-1	0	1	2
-2	1	$\cos \pi/8$	$\cos \pi/4$	$\cos 3\pi/8$	0
-1	$\cos \pi/8$	1	$\cos \pi/8$	$\cos \pi/4$	$\cos 3\pi/8$
0	$\cos \pi/4$	$\cos \pi/8$	1	$\cos \pi/8$	$\cos \pi/4$
1	$\cos 3\pi/8$	$\cos \pi/4$	$\cos \pi/8$	1	$\cos \pi/8$
2	0	$\cos 3\pi/8$	$\cos \pi/4$	$\cos \pi/8$	1

Table 6: Inter class weights used for computing inter annotated agreement



(a) Instructions for answer collection task



(b) Instructions for answer validation task

Figure 5: Instructions for MTurk tasks