PRAGMATICQA: A Dataset for Pragmatic Question Answering in Conversations

Peng Qi[*◁†] **Nina Du**[*△] **Christopher D. Manning**[△] **Jing Huang**[▷†]
[◁] AWS AI Labs
[△] Computer Science Department, Stanford University
[▷] Amazon Alexa AI
{pengqi, manning}@cs.stanford.edu

Abstract

Pragmatic reasoning about another speaker's unspoken intent and state of mind is crucial to efficient and effective human communication. It is virtually omnipresent in conversations between humans, e.g., when someone asks "do you have a minute?", instead of interpreting it literally as a query about your schedule, you understand that the speaker might have requests that take time, and respond accordingly. In this paper, we present PRAGMATICQA, the first large-scale open-domain question answering (QA) dataset featuring 6873 QA pairs that explores pragmatic reasoning in conversations over a diverse set of topics. We designed innovative crowdsourcing mechanisms for *interest-based* and *task-driven* data collection to address the common issue of incentive misalignment between crowdworkers and potential users. To compare computational models' capability at pragmatic reasoning, we also propose several quantitative metrics to evaluate question answering systems on PRAGMATICQA. We find that state-of-the-art systems still struggle to perform human-like pragmatic reasoning, and highlight their limitations for future research.

1 Introduction

Reasoning about interlocutors' unspoken intent or state of mind is a crucial feature of human communication, which allows us to convey ideas and exchange information more efficiently and effectively assuming that conversation participants are cooperative (Grice, 1975). For instance, when asked *"Is there water on Mars?"*, a friendly, knowledgeable person will not answer just *"Yes"*. Typically, the answerer would anticipate reasonable follow-up questions and/or identify the asker's theme of curiosity, and offer more details (see Figure 1 for an example). This capability of *pragmatic reasoning* is especially helpful when the asker is seeking information from an answerer that is more knowledgeable about the topic discussed, *e.g.*, in a teacher-student discussion, a user-database interaction (Kaplan, 1982), or a user-(virtual-)assistant conversation (Allen and Perrault, 1980).

Question: Is there water on Mars?
Literal, Direct Answer: Yes, there is water on Mars.

Potential follow-up question: *Where? In what form?*
Relevant knowledge: *Water has been found in 23 places in our Solar System. Turns out it isn't so parched.*

Pragmatic Answer: Yes, but only in the form of ice caps near its poles. In fact, Mars is just one of 23 places where we have found water in the Solar System!

Figure 1: An example of answering an information-seeking question literally vs. pragmatically by reasoning about the asker's unspoken information needs and potential relevant knowledge that might engage the asker.

Recent open-domain question answering (QA) datasets have placed an increasing emphasis on mimicking this information-seeking setting, but they still fall short at two crucial desiderata. First, most datasets mainly focus on evaluating systems' accuracy at finding the literal answer to a question, both in single-turn QA (Rajpurkar et al., 2016; Kwiatkowski et al., 2019) and multi-turn QA (Choi et al., 2018; Reddy et al., 2019). While this simplifies data collection and model evaluation, they cannot evaluate whether a QA system can understand or fulfill unspoken needs behind a question, which can be key to successful and engaging multi-turn interactions. Second, most of these datasets are crowd-sourced, which leaves them vulnerable to the problem of *incentive misalignment* between annotators and potential users (de Vries et al., 2020). This not only affects how or what questions are asked, but also how these questions are answered.

In this paper, we present PRAGMATICQA, a conversational open-domain question answering

[*] These authors contributed equally. PQ designed the collection mechanism and evaluation metrics, and wrote the paper; ND carried out most of the groundwork for collection/experiments/analyses. CM/JH offered project guidance and support, and helped discuss experiments and analysis.

[†] Work done prior to joining Amazon at JD AI Research.

Findings of the Association for Computational Linguistics: ACL 2023, pages 6175–6191
July 9-14, 2023 ©2023 Association for Computational Linguistics

Dataset	Open-domain	Multi-turn	Info-seeking	Extractive Rationale	Free-form Response	Topic-switching	Pragmatic Answers & Eval	Incentive-aligned
SQuAD (Rajpurkar et al., 2016)	✗	✗	✗	✓	✗	✗	✗	✗
Wizard of Wikipedia (Dinan et al., 2018)	✓	✓	✗	✗	✓	✗	✗	✗
Natural Questions (Kwiatkowski et al., 2019)	✓	✗	✓	✗	✗	✗	✗	⚠
QuAC (Choi et al., 2018)	✗	✓	✓	✓	✗	✗	✗	✗
CoQA (Reddy et al., 2019)	✗	✓	✗	✓	✓	✗	✗	✗
Curiosity (Rodriguez et al., 2020)	✗	✓	✓	✗	✓	✓	✗	✗
QReCC (Anantha et al., 2021)	✓	✓	⚠	✗	✓	⚠	✗	⚠
TopiOCQA (Adlakha et al., 2021)	✓	✓	✓	✗	✓	✓	✗	⚠
PragmatiCQA (this work)	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: Comparison between PragmatiCQA and previous work on key features. ✓, ✗, and ⚠ indicate a feature is present, absent, or partially represented in a given dataset. Please refer to the text for a more detailed discussion.

dataset that features conversations between humans that involve pragmatic reasoning, the first of its kind to the best of our knowledge. We also present various automated metrics to evaluate QA systems on answer accuracy, pragmatic reasoning, answer naturalness and faithfulness. Aside from pragmatic reasoning, PragmatiCQA is collected with incentive alignment as a primary design goal. To this end, we curate data with a focus on discussion topics that might share popular interest, and allow crowdworkers to choose topics of mutual interest to converse about instead of prescribing them. This allows crowdworkers to engage in conversations in a manner that closely mirrors a real user on topics they are genuinely curious about. Further, to encourage crowdworkers to explore the topic under discussion, we also design mechanisms where the question asker ("learner") can qualify as an answerer ("teacher") through learning and receive higher pay from the task. This not only allows us to collect high-quality conversational data with workers of varying amount of background knowledge on the same topic, but also aligns the diversity and quality of our data with crowdworkers' compensation. We finetune a Fusion-in-Decoder model on PragmatiCQA and find that our current models fails to recover >90% the pragmatic information that crowdworkers provided in the data.

To recap, our contributions in this paper are: 1) we propose PragmatiCQA, an open-domain conversational question answering (ConvQA) dataset featuring pragmatic answers and quantitative metrics to evaluate pragmatic reasoning in ConvQA; 2) we design a crowdsourcing framework for PragmatiCQA that alleviates the problem of incentive misalignment, which yields realistic, high-quality, and diverse data; 3) we analyze PragmatiCQA and show that it presents unique and important challenges to ConvQA systems today.[1]

2 Related Work

Our work is closely related to three topics, namely open-domain question answering (QA), conversational QA, and computational pragmatic reasoning, in which we review prior work in this section. We also highlight key features of PragmatiCQA and contrast it with previous work in Table 1.

Open-domain QA. The goal of open-domain QA is to answer questions from a large collection of unstructured knowledge (*e.g.*, text). SQuAD Open (Chen et al., 2017) is one of the most widely used datasets in this task, originally adapted from reading comprehension questions collected on Wikipedia passages (Rajpurkar et al., 2016). While it helps benchmark retrieval-based QA, the questions are often too context-dependent and ambiguous (*e.g.*, *"What day was the game played on?"*) or too unnaturally specific (*e.g.*, *"What park covers an area of 76 ha.?"*). In later work, Yang et al. (2018) expand open-domain QA to require multi-step reasoning which helped alleviate the former issue, but the latter remained unresolved.

TriviaQA (Joshi et al., 2017) and Natural Questions (Kwiatkowski et al., 2019) take two distinct approaches to improve incentive alignment in open-domain QA. While the former enlists trivia enthusiasts to author questions to reflect their interest, the latter takes questions typed into a search engine and answers them with crowdworkers on Wikipedia. However, as with other prior work, the question answerer is not incentivized to provide helpful answers that might address the asker's unspoken intent beyond a literal interpretation of the question.

Conversational QA. The growing interest in interactive natural language processing (NLP) sys-

[1]We release the data and code for the baseline at `https://github.com/qipeng/PragmatiCQA`.

tems has also driven the development of conversational QA (ConvQA) resources. Beyond reading comprehension QA tasks in conversations like QuAC (Choi et al., 2018) and CoQA (Reddy et al., 2019) and knowledge-enhanced chitchat like Wizard of Wikipedia (Dinan et al., 2018), there has also been growing interest in open-domain ConvQA tasks to closely imitate how virtual assistants operate in real life. QReCC (Anantha et al., 2021) and TopiOCQA (Adlakha et al., 2021) are two recent benchmarks in this direction. The former focuses on evaluating retrieval-based ConvQA systems on coreference and ellipsis resolution, and the latter is designed to train ConvQA systems to handle natural topic transitions. While both datasets attempt to simulate real-world information seeking by seeding conversations with questions from Natural Questions, these questions are assigned to crowd workers that are not necessarily interested in them, and thus the actual conversations might still fall short at closely modeling the conversation between a curious user and a helpful assistant.

Computational Pragmatic Reasoning. Since the publication of the Gricean cooperative principles between rational speakers (Grice, 1975), various frameworks have been proposed to characterize pragmatic reasoning in discourse understanding (Marcu, 1998) and in multi-agent communication, where the latter includes plan inference (Allen and Perrault, 1980), plan inference with discourse coherence (Asher and Lascarides, 1998, 2003), game theoretic analysis (Stevens et al., 2016), and rational speech acts (RSA; Frank and Goodman, 2012). When applied to textual responses of natural language interfaces, these techniques are often referred to as "over-answering" (Wahlster et al., 1983; Bersia et al., 1986) or "coorperative response generation" (Kaplan, 1982; Cheikes and Webber, 1989), which find their roots in database systems with natural language interfaces that serve users on knowledge-intensive tasks like question answering. We note, however, that most prior work focus on settings where all agents share all the referents involved for pragmatic reasoning, *e.g.*, the set of colors (Monroe et al., 2017), images (Cohn-Gordon et al., 2018), environments (Fried et al., 2018), and sometimes a finite set of utterances used to refer to them. Essentially, both speakers share the same information aside from the identity of the target referent (or the goal/plan) available only to the speaker, and computational approaches focus on efficient normalization over referents or utterances (Cohn-Gordon et al., 2018). In an information-seeking conversation, however, agents need to navigate the information asymmetry in their knowledge of potential referents, where shared common sense and pragmatic reasoning on the question answerer's part play an important role. We believe that PragmatiCQA provides a starting point and benchmark for the development of computational pragmatic reasoning approaches under information asymmetry with the full complexity of natural language.

3 PragmatiCQA: Pragmatic Question Answering in Conversations

In this section, we introduce how we crowdsource PragmatiCQA ranging from data source and processing to details about task design and how it helps align crowdworker interest with that of potential users. We conclude the section with evaluation metrics we propose to assess the pragmatic reasoning of QA systems, as well as statistics of the dataset.

3.1 Data Preparation

To engage crowdworkers in a teaching/learning conversation they are interested in, we select Fandom[2] as the source of our corpus for these conversations. Similar to Wikipedia, Fandom is a crowd-maintained web-based encyclopedia service on a wide variety of topics. Unlike Wikipedia, however, Fandom is largely organized around entertainment topics with content contributed by fans, where each topic is elaborated in a *community* of hundreds to thousands of webpages about each detail. As a result, Fandom is not only an ideal source of topics that might interest crowdworkers, but also offers a diverse set of relatively isolated topics to test models' few-shot or zero-shot generalization.

To select topics for data collection, we organize Fandom communities by their genres,[3] and select the most active communities as candidates for data collection (see Section 3.3 for more details). For each community, we manually choose a "central" page from which we follow hyperlinks up to three levels to scrape related topics. We remove navigation bars and sections from each page to limit the

[2]https://www.fandom.com/, also formerly known as Wikia. Fandom content uses CC-BY-SA license by default.

[3]We include eight genres in our corpus, namely Books, Games, Lifestyle, Comics, Music, Anime, TV, and Movies. We exclude Gaming, which is largely about e-sports tournament statistics, and filter out fan fiction due to its smaller audience.

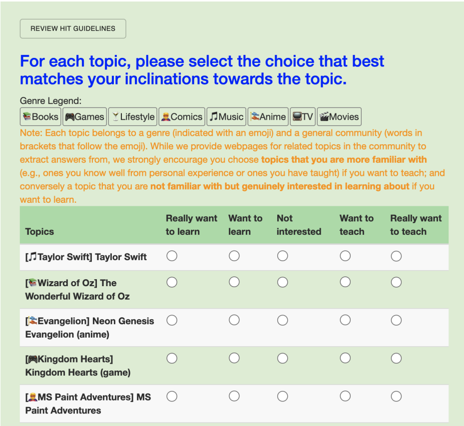

Figure 2: Example of a survey to identify topics of mutual interest between crowd workers.

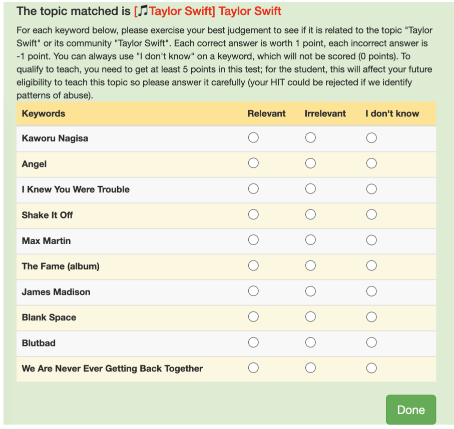

Figure 3: Example of a background knowledge test to assess crowd workers' background knowledge on a given topic.

scope of hyperlinks to the main body, which will serve as the reading material for crowdworkers to answer question from. We keep these hyperlinks in place for crowdworkers to navigate between webpages to find answers to questions. We discuss more details about the communities we selected for data collection in Appendix A.

3.2 Collecting Pragmatic Responses

In PRAGMATICQA, we pair crowdworkers to engage in a conversation about a topic drawn from Fandom. In each conversation, one crowdworker takes the role of the *teacher* and the other the *student*, and they are free to explore topics related to the central page of the topic.

Every conversation starts with a question from the student (*e.g.*, *"What is the Lord of the Rings?"*), which can be about a specific entity or event within the topic if it is not the first time a student is learning about it (*e.g.*, *"Who is Gandalf?"*). Then, the teacher attempts to answer the question with Fandom pages (that the student cannot access) in three steps. First, the teacher selects a collection of extracted spans that answers $\mathcal{A}_{\text{lit}}$ the question based on its literal interpretation (*e.g.*, *"The Lord of the Rings is an epic high fantasy novel written by J.R.R. Tolkien"*), or simply answer "Yes", "No", or "I don't know" if the pages do not contain the answer. This is similar to the extractive answers provided in previous conversational QA datasets (*e.g.* Reddy et al., 2019). Then, the teacher is tasked to select spans $\mathcal{A}_{\text{prag}}$ that might answer questions the student will ask next given the literal answer (*e.g.*, *"The story concerns peoples such as Hobbits, Elves, Men, Dwarves, Wizards, and Orcs (called goblins in The Hobbit), and centers on the Ring of Power made by the Dark Lord Sauron."*). Finally, to answer the question, the teacher is tasked to combine information from both span collections, and paraphrase it into a conversational response a.

The student is tasked to come up with a follow-up question once a response is received. As soon as this question is sent, the student is presented a survey about the response that was just received, regarding its naturalness, the quality of the span collections (not revealed until the survey), as well as the faithfulness of the final paraphrased response to these selected spans. This survey is answered concurrently with the teacher's answering of the follow-up question so that we minimize crowdworkers' wait time and keep the utterances as conversational as possible.

Once 6 rounds of QA pairs are reached, the crowdworkers can choose to leave the conversation at any time, and at the time they exit the task, we present both crowdworkers a short survey for feedback on the topic, the other crowdworker, and the task itself. We refer the reader to Appendix B for annotation guidelines and Appendix C for more details about the task interface.

3.3 Aligning Crowdworker Incentives with Potential Users

To align the interests of crowdworkers with potential users, we make several design choices to encourage crowdworkers to be genuinely curious about the discussion topic, to sufficiently explore it, and to produce good questions and responses in the task.

Incentive to learn. One common feature of crowdsourcing tasks that might lead to incentive misalignment is prescribed topics or roles, especially in a conversation, since they are unlikely to match crowdworkers personal interests or life experiences. To mitigate this issue, we begin by curating communities and topics from Fandom of popular interest. Specifically, we filter out Fandom communities with less than 100 content pages and 10 active users in the past 30 days, and rank them by the average edits per page as a proxy for user engagement. For each genre, we keep the top 30 communities as candidates for discussion topics for crowdworkers.

Beyond the curation of the pool of topics, we also provide crowdworkers means to indicate topics of mutual interest as they are paired to converse. At the beginning of each conversation, each crowdworker is shown a handful of potential topics for discussion, and asked to indicate their inclinations to teach or learn about each topic (see Figure 2). Once both crowdworkers have indicated their preferences, we will automatically select the topic that is the most compatible between the two crowdworkers for data collection, and assign teacher/student roles accordingly.

Incentive to explore. Although our topics are selected to match the student's curiosity and the teacher's self-proclaimed expertise, the topic alone does not guarantee that the student would explore knowledge about the topic as a curious user would, or that the teacher would be able to support them effectively. We incorporate two mechanisms to further align crowdworker incentives. First, both crowdworkers are paid more in rewards with each additional turn they finish beyond the minimum requirement, and the pay for each turn grows as the conversation persists longer. This encourages both crowdworkers to explore more topics within the conversation. Second, we design a simple *background knowledge test* for each crowdworker to gauge their readiness to teach a certain topic, inspired by Rodriguez et al. (2020). Similar to that work, we try to come up with multiple choice questions for each topic for crowdworkers to indicate their level of background in a given topic. However, since Fandom covers a large variety of diverse topics, we cannot easily come up with a fixed set of questions for each, like Rodriguez et al. (2020) did with geographic entities. We instead generate a list of popular page titles for each topic via personalized PageRank, and ask crowdworkers to answer which titles are related and which are unrelated from a list consisting of five relevant titles and five drawn from popular titles of other topics in the same genre (see Figure 3 for an example). This not only allows us to automatically generate these tests for any given topic, but also serves as an automatic qualification mechanism for crowdworkers to teach a topic, and incentivize students to explore the topic sufficiently should they want to teach the topic and get paid more per conversation.

3.4 Evaluation Metrics

Once a question answering model is built on PRAGMATICQA, we are interested in quantifying its performance with the data we have collected with the help of crowdworkers. Assume that a QA model produces predictions on all of the categories of information a crowdworker is asked to provide in PRAGMATICQA, namely a collection of literal answer spans $\hat{\mathcal{A}}_{\text{lit}}$ extracted from the webpages, a collection of pragmatic answer spans $\hat{\mathcal{A}}_{\text{prag}}$, and a final paraphrased answer $\hat{a}$.

Given these model predictions and their human-annotated counterparts, it is relatively straightforward to understand how accurate models are at answering the question based on its literal interpretation. We employ the standard F_1 metric for extractive question answering popularized by (Rajpurkar et al., 2016):

$$\text{F}_1^{\text{lit}} = \text{F}_1(\hat{\mathcal{A}}_{\text{lit}}, \mathcal{A}_{\text{lit}}), \tag{1}$$

which is effectively the same as the F_1 metric employed by previous extractive question answering datasets.

For PRAGMATICQA, we are further interested in how well models can learn to emulate the pragmatic behavior of human annotators. For this purpose, $\text{F}_1(\hat{\mathcal{A}}_{\text{prag}}, \mathcal{A}_{\text{prag}})$ would seem to be a good candidate. However, this metric does not account for the potential dependency between $\hat{\mathcal{A}}_{\text{prag}}$ and $\hat{\mathcal{A}}_{\text{lit}}$, between $\mathcal{A}_{\text{prag}}$ and $\mathcal{A}_{\text{lit}}$, as well as potential prediction errors. Ideally, we would like to capture the model's pragmatic reasoning *beyond* the information that is already in $\mathcal{A}_{\text{lit}}$ and assign a score of zero if no additional information is provided in overlap with $\mathcal{A}_{\text{prag}}$. In pathological cases where $\mathcal{A}_{\text{lit}} \cap \mathcal{A}_{\text{prag}} \neq \emptyset$, using $\text{F}_1(\hat{\mathcal{A}}_{\text{prag}}, \mathcal{A}_{\text{prag}})$ as the pragmatics metric allows predictions like $\hat{\mathcal{A}}_{\text{prag}} = \mathcal{A}_{\text{lit}}$ to receive non-zero scores, despite revealing no information that requires pragmatic reasoning. Comparing $\hat{\mathcal{A}}_{\text{prag}}$ to $\hat{\mathcal{A}}_{\text{lit}}$ is unlikely to be helpful, either, since it is possible to maximize their difference by setting $\hat{\mathcal{A}}_{\text{lit}} = \emptyset$. We therefore design the following metric to gauge the model's pragmatic reasoning against annotations:

$$\mathrm{F}_1^{\text{prag}} = \mathrm{F}_1(\hat{\mathcal{A}}_{\text{prag}} - \mathcal{A}_{\text{lit}}, \mathcal{A}_{\text{prag}} - \mathcal{A}_{\text{lit}}), \quad (2)$$

where $\mathcal{B} - \mathcal{A}$ removes all spans in $\mathcal{A}$ from spans in $\mathcal{B}$. It can be seen that this avoids the aforementioned pathological cases, and properly assigns a score of zero unless $\hat{\mathcal{A}}_{\text{prag}}$ contains information beyond $\mathcal{A}_{\text{lit}}$ that overlaps with $\mathcal{A}_{\text{prag}}$.

Last but not least, we are also interested in evaluating the final response $\hat{a}$. For this, we can apply reference-based evaluation metrics to compare it directly with a. Here, we apply the symmetric BARTScore (Yuan et al., 2021):

$$Q(\hat{a}) = \frac{\text{BARTScore}(\hat{a}, a) + \text{BARTScore}(a, \hat{a})}{2}. \quad (3)$$

Here, $\text{BARTScore}(x, y)$ uses a trained BART (Lewis et al., 2020) model on a text-to-text dataset (e.g., summarization or paraphrasing) to obtain the token-averaged conditional log likelihood of sequence y given sequence x as the input. BARTScore has been shown to exhibit better correlation with human judgement in a variety of tasks compared to prior model-based evaluation metrics. We use the symmetric formulation as a proxy for semantic equivalency, and we use the model fine-tuned on the CNN/DailyMail dataset.

3.5 Data Analysis

We analyze the conversations collected for PRAGMATICQA, present dataset statstics, and analyze the effect of our incentive alignment techniques.

As can be seen in Table 2, dialogues in PRAGMATICQA cover a broad set of topics, each of which contains an average of 7.8 question answer pairs. We split the data into train/dev/test sets with *disjoint* topics of discussion to minimize information leakage, so that the model's pragmatic reasoning capabilities can be evaluated on unseen topics to evaluate generalization. When studying questions in the dataset, we find that the teacher's response is usually significantly longer than that of the student's question (about 4x as long), which contains roughly the same number of words as literal and pragmatic answer spans combined. The literal and pragmatic answers have very little overlap overall (4.40 words per question), and typically come from 2.55 different HTML elements (usually different passages in the document). An average conversation in PRAGMATICQA extracts answer spans from 16.1 unique HTML elements from an average of 3.11 unique web pages, or a new page every 2.5 turns. This shows that PRAGMATICQA's setup encourages crowd workers to explore the topic of interest at a great depth, leading to natural topic shifts throughout the conversation.

PRAGMATICQA features a diverse set of questions (see Figure 4), which elicit a diverse set of literal answers, of which about 22.1% are "Yes/No/I don't know" answers with equal proportions. The rest of the literal answers contain a combination of short factoid answers and longer narrative answers

Split	QA	Dial.	Topics	Genres
Train	4027	526	34	8
Dev	1479	162	8	8
Test	1367	193	9	8
Total	6873	881	51	8

Component	Average Length
Q	8.34
A	31.37
$\mathcal{A}_{\text{lit}}$	13.47
$\mathcal{A}_{\text{prag}}$	22.90

Table 2: Statistics for different splits of PRAGMATICQA (top) and length statistics of questions and answers (bottom). Here, Q, A, $\mathcal{A}_{\text{lit}}$, and $\mathcal{A}_{\text{prag}}$ represent the the question, the paraphrased answer, all literal answer spans, and all pragmatic answer spans for each question, where average length is measured by the number of space-separated tokens.

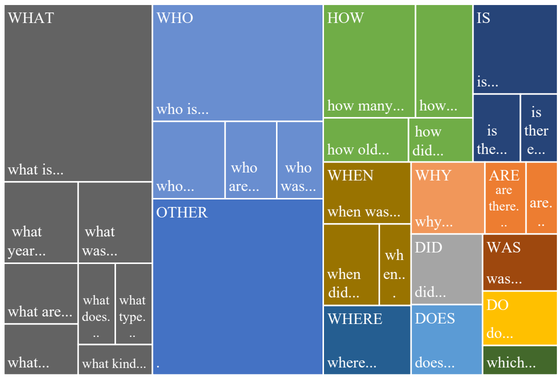

Figure 4: Question types featured in PRAGMATICQA. The area of each prefix corresponds to the proportion of questions that share the same prefix.

Q: When had the first Zelda game been released?
$\mathcal{A}_{\text{lit}}$: It came out as early as 1986 for the Famicom in Japan, and was later released in the western world, including Europe and the US in 1987. $\mathcal{A}_{\text{prag}}$: The Legend of Zelda is the first installment in the Zelda franchise, and its success allowed the development of sequels. In one or another way, nearly every title in the series is influenced by this game
A (a)**:** The Legend Of Zelda was first released as early as 1986 in Japan and later to the western world in 1987. It was the first installment in the Zelda franchise and its sucess allowed the development of sequels, with nearly every game in the series influenced by it!

Figure 5: An Example QA pair in PRAGMATICQA with literal and pragmatic answer spans. Taken from a conversation about "The Legend of Zelda".

		Non-empty		All
Incentive \ Stats		Secs/QA	QAs/Dial	QAs/Dial
Topic sel.	✓	341*	6.24	1.64
	✗	288*	5.52	1.36
BG test	✓	341	6.12	1.47*
	✗	301	6.12	2.14*

Table 3: Effect of incentive alignment techniques. "Topic sel." gives crowd workers freedom to choose topics to converse about, which incentivizes them to be curious and learn; "BG test" screens crowd workers that are not qualified to teach, which incentivizes them to fully explore each topic. ✓/✗ means the technique is enabled/disabled in A/B testing. * indicates results where the 95% confidence interval are disjoint. "Secs/QA" stands for the number of seconds crowd workers spend per QA pair, and "QAs/Dial" is the average number of QA pairs per dialogue.

(with quartile span lengths of 4, 12, and 21 tokens, respectively). Of the pragmatic answer spans, we find that 41% answer potential follow-up questions the Student worker might ask given the literal spans, 25% offer information from the web pages that helps sustain the conversation, and 22% do a bit of both. An actual example of PRAGMATICQA can be found in Figure 5.

We further study the effect of the incentive alignment techniques we presented in Section 3.3. Specifically, during data collection, we perform an A/B test for each technique, where we target 80% of completed conversations collected with each feature independently.[4] As can be seen in Table 3, when crowd workers are committed to converse on a topic ("non-empty", conversations with at least one QA pair), both techniques incentivize crowd workers to spend more time in the conversation, with a statistically significant gain observed on time per QA pair from allowing workers to select their topics of interest to discuss. Furthermore, we find that crowd workers are more likely to engage in longer conversations on a topic of their choosing, and spend more time to finish QA pairs when the Teacher worker is qualified through the background knowledge test. Finally, we find that the background knowledge test has a statistically significant filtering effect on conversations that could have taken place with an underqualified Teacher worker. That is, while non-empty conversations are qualitatively similar in length, the number of empty conversations as a result of enabling the background knowledge test significantly drives down the average QA pairs per conversation when they are considered. In contrast, the filtering effect for topic selection is much less pronounced, because the crowd worker has a wide variety of topics to choose from.

4 Experiments

4.1 Model and Setup

In our experiments to establish a baseline on PRAGMATICQA, we use a Fusion-in-Decoder (FiD) model (Izacard and Grave, 2021) with a dense passage retriever (DPR) (Karpukhin et al., 2020). DPR is a pretrained Transformer (Vaswani et al., 2017) retrieval model that helps us find correct passages from the Fandom corpus to answer questions, and FiD is a technique to combine top retrieved passages in a conditional generative model for efficient generation. We use BART-large (Lewis et al., 2020), a pretrained Transformer sequence-to-sequence model to generate answers.[5] We finetune the DPR question encoder and the BART model on the training set of PRAGMATICQA with the Adam optimizer (Kingma and Ba, 2015) with batch size of 4 and initial learning rate of 10^{-5} on two RTX 3090 GPUs, and select the model that achieves the best dev performance and stop training until the model fails to improve dev performance for 5 consecutive evaluations. The total training time is

[4] Specifically, we record how many successful conversations are conducted with/without each technique, and adjust the probability of enabling/disabling it in the next conversation accordingly. We also correct for recurring crowd worker(s) so that they cannot always skip configurations they do not like and introduce worker confounds in the results.

[5] We use the publicly available implementation and models from the Transformers (Wolf et al., 2020) library in our experiments and train them with ParlAI (Miller et al., 2017).

Top-k docs	dev					test				
	R@k	$\mathrm{R}^{\mathrm{All}}$@$k$	$\mathrm{F}_1^{\mathrm{lit}}$	$\mathrm{F}_1^{\mathrm{prag}}$	$Q(\hat{a})$	R@k	$\mathrm{R}^{\mathrm{All}}$@$k$	$\mathrm{F}_1^{\mathrm{lit}}$	$\mathrm{F}_1^{\mathrm{prag}}$	$Q(\hat{a})$
$k = 1$	1.59	1.49	9.85	5.23	-3.937	1.83	1.61	11.08	5.49	-3.931
$k = 5$	5.77	5.48	11.48	6.11	-3.754	4.90	4.24	11.92	5.73	-3.741
$k = 10$	8.12	7.51	11.25	5.99	-3.717	7.10	6.14	11.87	5.54	-3.655
$k = 20$	10.27	9.40	10.67	5.32	-3.666	9.62	8.34	11.94	5.27	-3.640
$k = 50$	14.52	13.12	10.27	5.00	-3.679	14.81	12.73	11.12	4.94	-3.636

Table 4: Performance of a Fusion-in-Decoder model on the PRAGMATICQA dataset when different number of documents are retrieved for each response generation. R@k stands for recall at k passages, and $\mathrm{R}^{\mathrm{All}}$@$k$ is the recall for the entire gold document collection at k passages.

about 4 hours. We report further training settings and hyperparameters in Appendix E.

The model is provided with unlimited conversation history during training and evaluation, and the output is formatted as follows: `Literal Span 1 </lit> ... </lit> Literal Span n </lit> Pragmatic Span 1 </prag> ... Pragmatic Span m </prag> Final Answer </s>`. We experiment with different top k passages used during evaluation, and report retrieval performance of DPR, $\mathrm{F}_1^{\mathrm{lit}}$ and $\mathrm{F}_1^{\mathrm{prag}}$ of the selected spans, and $Q(\hat{a})$ of the final answer.

4.2 Main Results

We report model performance by varying the number of top documents retrieved with DPR when each response is generated. As can be seen in Table 4, the recall of the gold context grows as expected with increasing k, which in turn also leads to improved final answer quality $Q(\hat{a})$ initially. We also find that the quality of literal answers ($\mathrm{F}_1^{\mathrm{lit}}$) selected improves slightly with more passages retrieved presumably with more gold contexts readily available in top retrieval results. However, we notice that $\mathrm{F}_1^{\mathrm{prag}}$ does not improve similarly. This cannot entirely be attributed to a lackluster retrieval performance, since the recall of the entire set of gold documents improves as the number of top documents increases, suggesting that predicting pragmatic spans is a more challenging task than predicting answer spans and fluent final answers.

4.3 Analysis

In this section, we analyze the prediction of the FiD model, highlight success and failure cases, and provide an empirical analysis of the proposed $\mathrm{F}_1^{\mathrm{prag}}$ metric. For these analysis, we use retrieve $k = 5$ top passages for each response, which has the best PRAGMATICQA performance on the dev set.

As can be seen in the examples in Figure 6, the trained model can exhibit several different failure modes. In the first example, we see that the model misinterpreted the Maze Runner book series as a film. While this is a common problem in single-

$\mathrm{F}_1^{\mathrm{prag}}$=0, $\mathrm{F}_1^{\mathrm{lit}}$=0, R@5=100.0, $Q(\hat{a})$=-3.151

Q1: What award did the Maze runner win?

Gold: ALA Best Fiction for Young Adults , The Maze Runner (2011) </lit> Young Reader's Choice Award , intermediate grades, The Maze Runner (2012) </prag> it won the ALA Best Fiction for Young Adults in 2011, and Young Reader's Choice Award in 2012. </s>

Pred:I don't know </lit> *The Maze Runner: The Scorch Trials* </prag> The film won the Academy Award for Best Visual Effects for its use of CGI. </s>

$\mathrm{F}_1^{\mathrm{prag}}$=6.67, $\mathrm{F}_1^{\mathrm{lit}}$=100.00, R@5=0.0, $Q(\hat{a})$=-3.131

Q2: Ok, Is there a signature color that Taylor has?

Gold: I don't know </lit> When Swift was in fourth grade, she won a national poetry contest with a three-page poem, "Monster in My Closet" </prag> I do not know of a signature color Taylor Swift has, but when she was in the 4th grade, she won a national poetry cont est with a 3 page poem called "Monster in My Closet". </s>

Pred: I don't know </lit> She is known for her bright red hair and bright blue eyes. </prag> I do not know. However, she has a bright red and blue color. </s>

$\mathrm{F}_1^{\mathrm{prag}}$=72.73, $\mathrm{F}_1^{\mathrm{lit}}$=8.16, R@5=100.0, $Q(\hat{a})$=-3.239

Q3: Where did Batman grow up?

Gold: Wayne Manor </lit> wealthy splendor and leads a happy and privileged existence until roughly the age of eight. </prag> Wayne Manor. There Batman lived in wealth and lead a happy and privileged life until he was eight years old. </s>

Pred: Born in Gotham City to Thomas Wayne ... and his wife Martha , Bruce Wayne grew up ... Wayne Manor estate ... their trusted butler Alfred J. "Alfred" Pennyworth . </lit> *He was brought up in Wayne Manor* and its wealthy splendor and leads a happy and privileged existence until roughly the age of twenty. </prag> Born in Gotham, Bruce Wayne is born in a mansion on Wayne Manor. He is brought up there by his butler, Alfred

Figure 6: Examples from PRAGMATICQA's dev set, the FiD model's predicted answers, and the metrics of the corresponding answers. For ease of reading, pragmatic spans are highlighted and roughly categorized as satisfying unspoken information needs, providing relevant knowledge, or *irrelevant to either*.

turn open-domain QA (Min et al., 2020), here, it should have been clear from the conversational context what the Student worker is referring to.

The second example exhibits a widely know issue with generative QA models known as factual inaccuracy or hallucination. Here, the model fabricated information about Taylor Swift's hair and eye colors, presumably triggered by the word "color" in the question. Note that in this case the model is also generating a paraphrased answer that is not consistent with the spans it generated.

Finally, the third example shows several issues. First, the predicted literal answer provides too much information that does not directly answer the question, unlike the succinct span that crowd workers annotated. Second, the predicted pragmatic span repeats information from the literal span. Third, the model also hallucinates when generating the pragmatic span, where instead of "the age of eight", the model generated "the age of twenty".

These examples suggest that current models still struggle with multiple facets of the task presented by PRAGMATICQA: retrieval accuracy, factually grounded generation, generation consistency, entity disambiguation, as well as the ability to retrieve pragmatically useful information to present. We do note, however, that the proposed F_1^{prag} properly awards models for surfacing information that is not in the gold literal spans but in the gold pragmatic spans regardless of what the predicted literal answer is, effectively decoupling the evaluation of the two. We find that the full suite of proposed metrics, when used together, can holistically evaluate the answer quality and pragmatic reasoning strength of the conversational QA model.

5 Discussion: Evaluation on PRAGMATICQA

While the proposed evaluation metrics are useful to provide quality estimates of the provided answers, especially when it comes to how well the prediction matches the annotators' pragmatic behavior, we acknowledge that this is far from a complete set of evaluations desirable for real-world systems developed on PRAGMATICQA to be useful.

First, the evaluation metrics presented do not evaluate the final system response on its factuality or faithfulness to the spans selected, which is crucial for real-world systems. Liu et al. (2023) recently report that publicly available generative search engines are still far from satisfactory on this front, which we speculate will be more challenging for pragmatic responses such as those in PRAGMATICQA.

Second, unlike the literal answer, the definition of pragmatic responses in an information-seeking conversation is open-ended and more subjective in nature. In this paper, we explored categorizing these into two broad categories, answers to address potentially unspoken information needs, and potential relevant knowledge that can be helpful, but this is far from comprehensive, since good pragmatic responses could involve clarification questions that are not covered by PRAGMATICQA. Even within these categories, we see that at a given point in the conversation, there are typically more than one follow-up questions to be asked given the literal response; relevant knowledge is only more diverse. While a high F_1^{prag} score can approximate a sufficient condition for a pragmatic natural language system, it might be far from necessary due to the potential existence of multiple good answers.

Both of these suggest that additional evaluation metrics are necessary for PRAGMATICQA, which we leave to future work. We believe, given these observations, that model-based evaluation will become crucial in the pursuit of better evaluation methods on PRAGMATICQA, where our dataset will provide the resource to help kickstart the exploration. In the meantime, we also believe that the metrics presented in this paper can still serve as good proxies for evaluating model's pragmatic behavior until more powerful evaluation methods are available.

6 Conclusion

We presented PRAGMATICQA, the first open-domain conversational question answering dataset featuring pragmatic answers and quantitative metrics to evaluate pragmatic reasoning in conversational QA. PRAGMATICQA is collected with innovative crowdsourcing techniques, including techniques that better align crowd worker incentive with eventual users of ConvQA systems that improves crowd worker engagement and data quality. Finally, we show in our experiments that questions in PRAGMATICQA present unique and important challenges to ConvQA systems today, and open new research directions for investigation.

7 Limitations

PRAGMATICQA is collected via crowdsourcing on English-language material from Fandom.com, where community-maintained wiki pages are used as reading materials and basis for answering questions. Therefore, it cannot be guaranteed that the excerpts from Fandom will be factually correct or stay unchanged over time, and in turn the answers in PRAGMATICQA are also not factually verified. Furthermore, techniques or models developed on PRAGMATICQA might not be generally applicable to non-English languages or non-entertainment topics without further adjustment or evaluation.

More importantly, the crowd workers that participated in PRAGMATICQA are geographically limited to primarily English-speaking countries, and therefore might not represent typical pragmatic reasoning behaviors of people that speak different first languages or come from different cultural backgrounds. Therefore, it should not be treated as a universal standard for pragmatic reasoning in information-seeking conversations, but rather a single reference point.

Acknowledgments

The authors would like to thank the anonymous reviewers for the discussion and suggestions. This research was supported in part by the SAIL-JD Research Initiative.

References

Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. 2021. TopiOCQA: Open-domain conversational question answeringwith topic switching. *arXiv preprint arXiv:2110.00768*.

James F Allen and C Raymond Perrault. 1980. Analyzing intention in utterances. *Artificial intelligence*, 15(3):143–178.

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. Open-domain question answering goes conversational via question rewriting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 520–534, Online. Association for Computational Linguistics.

Nicholas Asher and Alex Lascarides. 1998. Questions in dialogue. *Linguistics and Philosophy*, 23(3):237–309.

Nicholas Asher and Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

Nicoletta Bersia, Barbara Di Eugenio, Leonardo Lesmo, and Pietro Torasso. 1986. The overanswering mechanism in the Fido system. In *Cybernetics and Systems '86: Proceedings of the Eighth European Meeting on Cybernetics and Systems Research, organized by the Austrian Society for Cybernetic Studies, held at the University of Vienna*, pages 823–830. Springer.

Brant A. Cheikes and Bonnie L. Webber. 1989. Elements of a computational model of cooperative response generation. In *Speech and Natural Language: Proceedings of a Workshop Held at Philadelphia, Pennsylvania, February 21-23, 1989*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Reuben Cohn-Gordon, Noah Goodman, and Christopher Potts. 2018. Pragmatically informative image captioning with character-level inference. 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 439–443, New Orleans, Louisiana. Association for Computational Linguistics.

Harm de Vries, Dzmitry Bahdanau, and Christopher Manning. 2020. Towards ecologically valid research on language user interfaces. *arXiv preprint arXiv:2007.14435*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

Michael C Frank and Noah D Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. 2018. Speaker-follower models for vision-and-language navigation. *Advances in Neural Information Processing Systems*, 31.

H. P. Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Speech Acts: Syntax and Semantics Volume 3*, pages 41–58. Academic Press, New York.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

S Jerrold Kaplan. 1982. Cooperative responses from a portable natural language query system. *Artificial Intelligence*, 19(2):165–187.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Nelson F Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. *arXiv preprint arXiv:2304.09848*.

Daniel Marcu. 1998. A surface-based approach to identifying discourse markers and elementary textual units in unrestricted texts. In *Discourse Relations and Discourse Markers*.

Alexander Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. 2017. ParlAI: A dialog research software platform. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 79–84, Copenhagen, Denmark. Association for Computational Linguistics.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.

Will Monroe, Robert X.D. Hawkins, Noah D. Goodman, and Christopher Potts. 2017. Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*, 5:325–338.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Pedro Rodriguez, Paul Crook, Seungwhan Moon, and Zhiguang Wang. 2020. Information seeking in the spirit of learning: A dataset for conversational curiosity. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8153–8172, Online. Association for Computational Linguistics.

Jon Scott Stevens, Anton Benz, Sebastian Reuße, and Ralf Klabunde. 2016. Pragmatic question answering: A game-theoretic approach. *Data & Knowledge Engineering*, 106:52–69.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wolfgang Wahlster, Heinz Marburger, Anthony Jameson, and Stephan Busemann. 1983. Over-answering yes-no questions: Extended responses in a nl interface to a vision system. In *Proceedings of the Eighth international joint conference on Artificial intelligence-Volume 2*, pages 643–646.

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.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34.

Genre	Communities	Examples
Anime	14	Soul Eater, One Piece, Studio Ghibli
Books	12	H. P. Lovecraft, Wizard of Oz, The Maze Runner
Comics	8	Sonic the Hedgehog, Batman, Peanuts Comics
Games	8	Halo, Fallout, The Legend of Zelda
Lifestyle	6	Olympics, The Formula 1, LEGO
Movies	10	Pixar, Harry Potter, The Matrix
Music	7	Lady Gaga, 'Cats' Musical, Taylor Swift
TV	8	Doom Patrol, Game of Thrones, Doctor Who

Table 5: Communities used in PRAGMATICQA collection.

A Fandom Communities Used in Collection of PRAGMATICQA

We collect data for PRAGMATICQA on eight genres of Fandom communities, and ensure that the coverage for genres is roughly even. Table 5 contains the number of communities used in each genre and example communities. We select up to 1,000 pages from each community by following hyperlinks from a hand-chosen landing page to up to three levels. The resulting average community used during data collection contains about 390 web pages each, which results in 401,042 DPR passages when processed and converted into plaintext. Each HTML element is marked with a unique UUID key to record span start and end during data collection, so that PRAGMATICQA can provide strong supervision for answer spans rather than relying on distant supervision post hoc.

B Guidelines for Crowdworkers

Please see Figure 7 for the guidelines we use for our crowdsourcing task.

C Crowdworker Interface

Please see Figure 8 for an example of our crowdsourcing interface. Our interface is built on the Mephisto toolkit[6] and ParlAI (Miller et al., 2017).

Both interfaces consist of a chat window with the full chat history that allows crowd workers to type in questions and answers on the right, and a side pane on the left that displays functional elements. For the teacher, the side pane contains instructions and controls to select spans from the web page below to serve as the literal and pragmatic answers, as well as an embedded web page with hyperlinks and back and forward controls to emulate a basic browser. For the student, the side pane contains basic task information and instructions for most parts of the task, and when an answer is available from the teacher, the student is tasked to rate it on answer quality, how well it addresses the student's unspoken information needs, as well as how faithful the final paraphrased answer is to the spans selected from web pages.

D Analyzing Question Types Featured in PRAGMATICQA

To determine the question type, we first locate WH-words (what, when, where, who, whom, which, whose, why, how) in the question. When that fails, we attempt to locate auxiliary verbs (is, are, was, were, did, do, does). From these words, we count up to three words to the right and summarize the salient patterns. When neither a WH-word or an auxiliary verb can be found, we categorize the question as "OTHER", which can include imperatives like "Tell me more about ...".

E Additional Hyperparameters

During training, we truncate input texts to at most 512 tokens and the concatenated output to 128 tokens for efficiency. We retrieve top 5 documents for FiD training. The model is evaluated on the dev set every 0.25 epochs during training, and learning rate is halved every time dev perplexity fails to improve. We stop training if dev perplexity does not improve for five consecutive evaluations, and select the model that achieves the best dev perplexity during training.

[6]https://github.com/facebookresearch/Mephisto

Teaching and Learning in a Conversation

Background

In this task, you will be invited to chat with another crowd worker to teach them or learn from them something you are both interested in.

The goal of this task is to collect data to teach computer assistants (think Siri, Alexa) to answer our questions engagingly and helpfully. Specifically, we would like to teach computers to answer questions by addressing the unspoken **intent** behind them and providing **helpful leads** when appropriate.

What do we mean by **intent** and **helpful leads**? Consider asking your friend who's knowledgeable about astronomy: "Is there water on Mars?"

Your friend's answer is probably more informative than the "robotic" answer, "Yes, there is water on Mars.", which is what today's computer assistants tend to offer, as they tend to interpret questions literally.

In contrast, sensing your desire to learn a bit about water on Mars if it does exist, your friend would probably say something like "Yes, but only in the form of ice caps at its poles." This can be seen as them anticipating your relatively predictable follow-up question "In what form?" and addressing both questions in a single response, which is part of the unspoken intent of the asker. In this task, we define "**intent**" as unspoken needs of information that can be reasonably inferred after the question is answered literally.

Aside from this, your friend might also know about water on other planets in the Solar System, and mention that in their response. Although not directly answering your original question, it would help to prompt you to engage and explore more in the conversation and learn from their knowledge. In this task, we call this "**helpful leads**", which we define as extra information from the answerer's knowledge that would help the asker explore beyond their original question.

Note a good answer might both address the asker's unspoken intent and offer helpful leads to engage the asker.

In the HIT, we will assign Turkers in the role of either a **student** or a **teacher**, where the student's task is to ask inquisitive and relevant questions about a topic they are interested in learning about but have limited knowledge of, and the teacher's task is to help us answer the question both literally and helpfully so we can quantify the difference between the two.

Task workflow

1. If you are working on this task for the first time, you will be asked to complete a qualifying task that familiarizes you with the idea of literal answers and helpful answers.

2. Once you pass the qualification, we will pair you with another Turker to engage in a teaching/learning conversation. Before the conversation starts, we will first ask each of you your interest in teaching or learning about a set of 5–15 topics. Your mutual interest will determine the topic of the conversation, as well as the teacher/learner role assigned.

3. Once a topic is chosen, the learner can start by asking the first question.

4. Given each question, the teacher will follow instructions to help us answer the question literally first, then furnish helpful information to engage in a more friendly conversation.

5. Given each response from the teacher, the student will be asked to rate it on different aspects following the instructions on our task interface.

6. You would need to each finish 6 rounds of conversation (plus rating for the student) to complete a conversation in this task.

7. At the end of the conversation, the teacher will be asked to rate the learner on several aspects, to determine if they are engaged and asking meaningful questions.

[... Task reward section omitted ...]

Disclosure and Consent

By participating in this task, you acknowledge and give explicit consent that we record your MTurk ID (but no personal identifiable information otherwise) to improve our ability to match you with topics you might be interested in teaching or learning about, as well as record your qualification status to perform the task and to teach a particular topic assigned to you. You further give us permission to release an anonymized version of your MTurk ID (that cannot be traced back to you or your actual MTurk ID) along with the data we collected, to help future researchers study how different people approach this task, as well as how your knowledge of a certain topic might have evolved by participating in this task.

Figure 7: Guidelines for the crowdsourcing task for PRAGMATICQA.

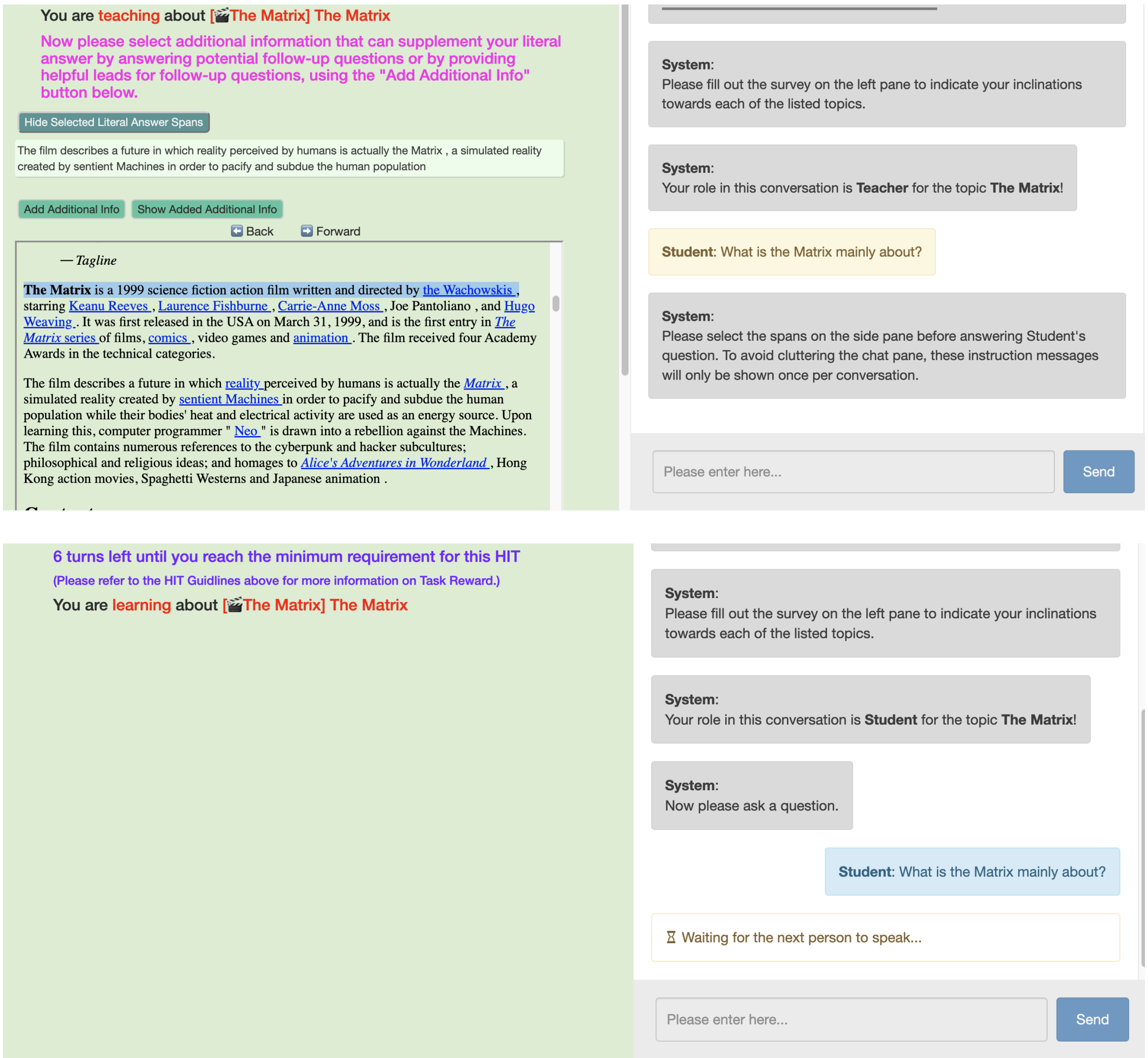

Figure 8: Crowdsourcing interface for PRAGMATICQA's teacher worker (top) and student worker (bottom).

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
6

☑ A2. Did you discuss any potential risks of your work?
6

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

3,4

☑ B1. Did you cite the creators of artifacts you used?
3,4

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
3

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
3

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. The data is based on public, community maintained data with caveats discussed in the Limitations section.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
3, Appendix A

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
3.5

C ☑ Did you run computational experiments?

4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
4

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
4, Appendix E

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
4

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
3

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
3, Appendix B, Appendix C

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
3, 6

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Appendix B

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
6