CREPE: Open-Domain Question Answering with False Presuppositions

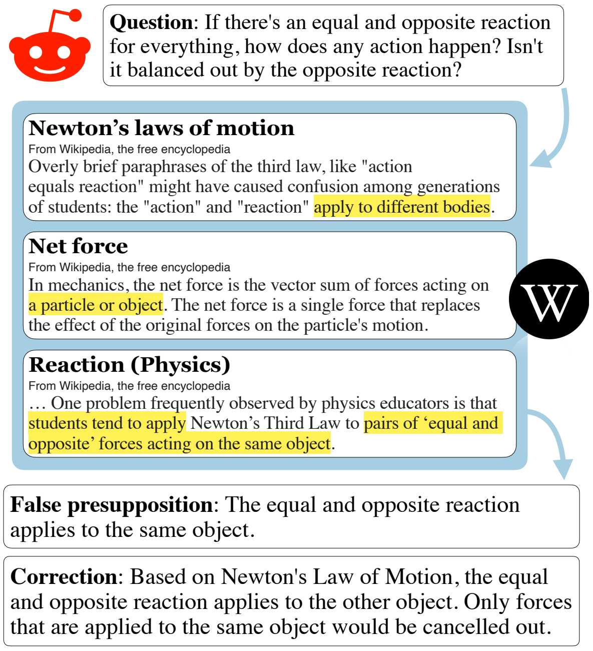
Xinyan Velocity Yu[†] **Sewon Min**[†] **Luke Zettlemoyer**[†] **Hannaneh Hajishirzi**[†,‡]
[†]University of Washington [‡]Allen Institute for Artificial Intelligence
`{xyu530,sewon,lsz,hannaneh}@cs.washington.edu`

Abstract

When asking about unfamiliar topics, information seeking users often pose questions with false presuppositions. Most existing question answering (QA) datasets, in contrast, assume all questions have well defined answers. We introduce CREPE, a QA dataset containing a natural distribution of presupposition failures from online information-seeking forums. We find that 25% of questions contain false presuppositions, and provide annotations for these presuppositions and their corrections. Through extensive baseline experiments, we show that adaptations of existing open-domain QA models can find presuppositions moderately well, but struggle when predicting whether a presupposition is factually correct. This is in large part due to difficulty in retrieving relevant evidence passages from a large text corpus. CREPE provides a benchmark to study question answering in the wild, and our analyses provide avenues for future work in better modeling and further studying the task.[1]

1 Introduction

When an information-seeking user poses a question about the topic they are unfamiliar with, they can often introduce false presuppositions (Kaplan, 1978; Duží and Číhalová, 2015) which are assumed but not directly stated. For instance, the question in Figure 1 incorrectly presupposes that the equal and opposite reactions in Newton's law apply to the same object. Although such a question is unanswerable, we might still hope to identify the confusion and explain it to the user. This functionality goes well beyond prior open-domain QA task formulations, which focus on questions with a valid direct answer (Rajpurkar et al., 2016; Kwiatkowski et al., 2019) or that are unanswerable from lack of evidence (Rajpurkar et al., 2018; Choi et al., 2018; Asai and Choi, 2021). While recent work studies unverifiable presuppositions in reading-comprehension-style questions given evidence context (Kim et al., 2021), there has been no work that identifies and corrects a presupposition that is false based on *global, factoid* knowledge.



Question: If there's an equal and opposite reaction for everything, how does any action happen? Isn't it balanced out by the opposite reaction?

Newton's laws of motion
From Wikipedia, the free encyclopedia
Overly brief paraphrases of the third law, like "action equals reaction" might have caused confusion among generations of students: the "action" and "reaction" apply to different bodies.

Net force
From Wikipedia, the free encyclopedia
In mechanics, the net force is the vector sum of forces acting on a particle or object. The net force is a single force that replaces the effect of the original forces on the particle's motion.

Reaction (Physics)
From Wikipedia, the free encyclopedia
… One problem frequently observed by physics educators is that students tend to apply Newton's Third Law to pairs of 'equal and opposite' forces acting on the same object.

False presupposition: The equal and opposite reaction applies to the same object.

Correction: Based on Newton's Law of Motion, the equal and opposite reaction applies to the other object. Only forces that are applied to the same object would be cancelled out.

Figure 1: An example question written by an online user that contains a false presupposition. The model is required to (a) identify the presupposition made in the question that is false based on world knowledge, and (b) write the correction. We also show three evidence paragraphs from the English Wikipedia.

In this paper, we introduce **CREPE** (**Cor**R**E**ction of **Pr**E**supposition), a new dataset consisting of 8,400 Reddit questions with (1) whether there is any false presuppositions, and (2) if any, the presupposition and its correction written. We find 25% of questions on Reddit (Fan et al., 2019) include false presuppositions, where the best response is to provide a correction of the presuppositions.

While the CREPE annotation task is challenging due to the need for extensive background knowl-

---

[1] The data, baselines, and the evaluation script are available at `github.com/velocityCavalry/CREPE`.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 10457–10480
July 9-14, 2023 ©2023 Association for Computational Linguistics

edge and inherent debatability, we leverage the *most upvoted comments* written by community users to efficiently annotate the data. The most upvoted comments are likely to be factually correct, and typically also identify and correct any false presuppositions made in the question. By designing an annotation pipeline using these comments, we were able to collect high-quality data at a relatively low cost. Our data analysis (Table 2) shows that the types of false presuppositions are diverse, ranging from relatively explicit presuppositions (e.g., false clauses or false predicate) to subtle, nuanced presuppositions (e.g., false causality or false existential presuppositions).

We define two tracks with varying levels of difficulty, and introduce models to set baseline performance levels for each. A model is given either the question only (the main track) or the question and the comment (the GOLD-COMMENT track), and is supposed to perform two subtasks: identification of whether or not there is a false presupposition (the **detection** subtask) and generation of presuppositions and their corrections (the **writing** subtask). For the writing subtask, we propose a systematic human evaluation scheme based on Celikyilmaz et al. (2020) that considers fluency, correctness (precision of the information), adequacy (recall of the information) and consistency.

We include a range of baselines, including a question-only model, a nearest-neighbor model, and a competitive model based on the state-of-the-art passage retrieval (Krishna et al., 2021) and pretrained language models (Liu et al., 2019; Raffel et al., 2020). Results and analyses indicate that (1) retrieval is very challenging since simply retrieving passages in the topic of the question is not enough; (2) models do moderately well in identifying explicit false presuppositions, and (3) models struggle with identifying implicit presuppositions and explaining how and why the presupposition is false. We also discuss open problems, such as an inherent ambiguity in the validity of presuppositions and inconsistency between different websites.

2 Background

2.1 Question Answering

There has been significant work on question answering, where the model receives a natural language, open-domain question and is required to return a short, concise answer (Voorhees and Tice, 2000; Lee et al., 2019). Most work focuses on questions that have a short text span as a correct answer. Other work studies unanswerable questions, but they study questions that are either intentionally written to be unanswerable (Rajpurkar et al., 2018) or where there is a lack of evidence to find the answer (Choi et al., 2018; Asai and Choi, 2021). More recently, Kim et al. (2021) studies unverifiable presuppositions in questions under the given context, but using questions from Kwiatkowski et al. (2019) whose presuppositions are mostly not false based on *global* knowledge.[2]

In this work, we focus on open-domain questions with presuppositions that are false based on global knowledge, and argue that an adequate response to them is a correction of these presuppositions. We show that false presuppositions in questions are **prevalent**. 25% of questions contain false presuppositions in the domain of online forums (for which we collect annotations), research papers (Dasigi et al., 2021), scientific reviews (Kang et al., 2018), and social media (Sap et al., 2020).

2.2 Presupposition

Under a *pragmatic* point of view, a presupposition is a condition that a speaker would normally expect to hold in the common ground between discourse participants when that sentence is uttered (Beaver et al., 2014; Stalnaker et al., 1977). Unlike semantic presuppositions (Strawson, 1950), pragmatic presuppositions cannot easily be traced to specific words or phrases, but rather depend on the context and the expectations of the discourse participants (Potts, 2015). A key property of pragmatic presuppositions is that they are *backgrounded*—a pragmatic property of being a meaning that the speaker presumes to be mutual public knowledge. While whether or not one is a presupposition is inherently debatable, we define a false presupposition based on the most voted comment in the online forum, as we discuss further in Section 3.2.

False presuppositions in *questions* have been discussed in the linguistic literature (Kaplan, 1978; Duží and Číhalová, 2015). False presuppositions in the question make a question infelicitous because there is no direct answer. Kaplan (1978) claims that an adequate and unambiguous answer to such a question is a negated presupposition, referred to as corrective indirect response. We follow them in providing the negated presupposition as the response

---

[2] Less than 5% of questions from Kwiatkowski et al. (2019) contains false presupposition under our definition, likely because their questions are aggressively filtered.

to the question, and build the first benchmark based on questions written by information-seeking users.

3 Dataset: CREPE

Our task requires the model to be given a question q and to (a) identify whether q has any false presuppositions, and (b) if yes, generate the false presupposition as well as the correction.

We have three criteria in constructing the dataset:

C1. Naturalness of the questions: whether the questions in the data are written by real, information-seeking users.

C2. Validity of the presupposition: whether the identified presupposition is highly likely made by the question writer.

C3. Correctness and adequacy of the information: whether the information provided in the correction is factually correct, and adequate to convince the question writer.

We first describe the source of questions (Section 3.1) which address **C1**. We then describe the process of annotation (Section 3.2) that addresses **C2** and **C3**. Finally, we present detailed, qualitative analysis of the data (Section 3.3). The formal definition of the task and metrics are in Section 4.

3.1 Data Source

Our highest priority is to study false presuppositions that *naturally* occur from information-seeking users. While it is significantly easier to manually write questions that would have false presuppositions, we think these questions will significantly be different from naturally occurring questions.

Following Fan et al. (2019), we use questions posted on the ELI5 subreddit.[3] We made a few modifications to the procedure that Fan et al. (2019) took in order to improve the data quality. We first filter questions and comments based on upvotes with a higher threshold. We then split the training, the development and the test data based on the time of the posting: questions on the training set are posted in 2011–2018, questions on the development set are posted in Jan–Jun of 2019, and questions on the test set are posted in Jul–Dec of 2019. Appendix A provides more details.

Krishna et al. (2021) raised a concern that a significant amount of test are duplicates of those on the training set. We provide a detailed analysis in Appendix A. In summary, we think (1) the amount of duplicated (or paraphrased) questions is significantly less than their estimate with respect to underlying presuppositions in the questions, and (2) even if there are paraphrased questions, data split based on the time frame is justified based on the real-world scenario.[4]

3.2 Data Annotation

Meeting the criteria **C2** and **C3** can be very difficult for the following reasons:

- For **C2**: The validity of presupposition is inherently debatable and largely depends on the background of individuals (Section 2.2).[5]
- For **C3**: The open-domain nature of the task requires the search of world knowledge on the web, which is extremely expensive and may not be exhaustive enough despite the best efforts made by annotators, as discussed in Kwiatkowski et al. (2019); Min et al. (2020).

In this work, we make use of the **most upvoted comments** written by community users. The comment, often written by domain experts, provides a response to the question in the ELI5 subreddit and has been used as a credible source in prior work (Fan et al., 2019). If the comment identifying a false presupposition has the most upvotes, it is likely that the presupposition is valid (made by a question writer) based on the background context shared by community users, thus satisfying **C2**. Moreover, the comment (1) is highly likely to contain information that is correct and adequate (satisfying **C3**), and (2) removes the need for exhaustively searching over the web (reducing the annotation cost).

Annotation task. Annotators are given a pair of the question and the most voted comment, and perform the following steps.

1. Filter out questions that are subjective, are uninformative, or rely on personal experience.
2. Judge whether there is a false presupposition in the question, identified by the comment.
3. If there is a false presupposition, write the presupposition and a correction as a concise, declarative sentence.

---

[3] `www.reddit.com/r/explainlikeimfive`

[4] We think having similar questions is an inherent property of questions on the web, and the model should be allowed to take whichever approach that is plausible, including the nearest neighbor approach (Lewis et al., 2021).

[5] This is also the case in previous work—for instance, the data annotated by experts in formal semantics and pragmatics can have low agreement (Jeretic et al., 2020).

Data split	# Questions		# Tokens				Posting time
	Tot	w/ FP	Q	PS	CR	CM	
Training	3,462	907 (26.2%)	15.6	10.3	16.5	95.2	2018
Development	2,000	544 (27.2%)	16.1	10.3	15.6	91.0	Jan–Jun 2019
Test	3,004	751 (25.0%)	16.4	11.8	16.8	92.5	Jul–Dec 2019
Unlabeled training	196,385	-	15.7	-	-	96.6	2011–2018
Total (labeled only)	8,466	2,202 (26.0%)	16.0	10.8	16.5	93.3	
Total (labeled+unlabeled)	204,851	-	15.7	-	-	96.5	

Table 1: Data statistics. *# Questions* indicate the number of questions in total and with false presupposition. *# Token* indicate the number of tokens (based on the whitespace tokenization) in the question (Q), the presupposition (PS), the correction (CR), and the comment (CM). Note that PS and CR are those to be written by the model; CM is given to the model in the GOLD-COMMENT track. Details of tracks are provided in Section 4.

Annotation pipeline. We maintain a pool of qualified annotators who passed our qualification task. We assign two annotators per question, where each annotators independently annotate the question. We filter out questions if either of the annotators mark them as such. If the annotators agreed to on the label (whether or not there is a false presupposition), their label as well as their writings are taken as gold references. When they disagreed, we assign a third annotator and take a majority vote over three workers. The percentage agreement in the initial stage of the annotation is 75%, and the Fleiss' kappa is 43%, indicating moderate agreement.[6] We find that the disagreements are mainly due to inherent ambiguities of the task due to the different interpretations of the question or the comment, or difference in individual background knowledge (discussion in Section 3.3).

More details in instructions and quality control are provided in Appendix B.

3.3 Data Analysis

The data statistics are provided in Table 1. We find that over 25% of questions posted on Reddit includes false presuppositions.[7]

Categorization of false presuppositions. We randomly sample 50 questions with false presuppositions and categorize them in Table 2. The five most frequent types of presuppositions include:

- **False clauses** are those where `FP` is made as a clause in `Q`, e.g., "the water has to be 100 to become steam" in Table 2.
- **False predicate** are those where the predicate in `Q` is false, e.g., "current is stored in power plants".
- **False properties** are those where certain properties or attributes are presupposed in `Q`, e.g., "cement blocks are too strong so that people who punch them are likely to break their hand." They are often very implicit and may be deniable; however, the question writer would not have asked the question if they have not made the presupposition.
- **False (causal) relationship between facts** are those where `Q` makes a (causal) relationship between facts that are false. This is another implicit type of presuppositions, but again, the question writer would not have asked this question if they have not made such a presupposition. A large portion of questions involve scientific phenomenon on which the question writer has misunderstanding, e.g., in the example in Table 2, the question writer had a misconception about Newton's Third Law of Motion.
- **False existential presupposition** indicates that `Q` includes an existential presupposition, one type of semantic presuppositions, that is false. For instance, the example `Q` in Table 2 presupposes an unused space in a hard disk, and `C` says there is no unused space.

Triggers in the comment. We analyze how comments point out the falsehood of the presupposition made in the question on the same set of 50 samples on the development data. In 68% of times, comments include specific lexical cues which we call *triggers*. 70% of such triggers are negations. Other word-level triggers include "actually", "just", "rare", "really" and "though". Sentence-level triggers include "You are mistaken", "You've got some

---

[6] Kappa is a chance-corrected coefficient and is correlated to the skewness of data distribution.

[7] Xu et al. (2022), who also used the ELI5 subreddit, identified 10% of questions have false presuppositions (rejected presuppositions) and excluded them in their data. Their estimate is significantly lower than ours because they did not include partial rejection while we do.

False clauses (14%)

`Q` If water has to be 100 to become steam, how come you don't get heavily burned in saunas?
`C` What we often call steam is just water vapor that has started to become visible due to different temperatures of the water vs air. It can exist at many temperatures. (...)
`FP` Water can steam below 100 degrees.

False properties (22%)

`Q` How do martial artists who karate chop or punch a cement block not break their hand? `C` It's a trick, the blocks are not very strong, and they are being punched or kicked in their weakest points. `FP` Chops or cement blocks are strong.
`Q` How does your phone tell the difference between a step and random movement? `C` You might be disappointed by this answer, but most of the time, you're not moving your phone (...) when you walk. `FP` A random movement is detectable by a phone.

False existential presupposition (6%)

`Q` What uses the space on a hard disk that we're unable to use? For example in a 1TB hard disk, we get about 930GB of usable memory, what happens to the other 70GB?
`C` There are TB (terabyte) and TiB (tebibyte). the "ra" ones are using multiples of 1000. the "bi" ones are using multiplies of 1024. I will do some math for you: 1 TB=1000^4B = (...) = 0.93 TiB. There goes your 70 GiB.
`FP` In a 1TB hard disk, 70GB is unusable.

False predicate (30%)

`Q` How exactly is current stored in power plants?
`C` It's not being stored at all. The power grid is a carefully balanced dance of supply and demand. (...)
`FP` Current is stored in power plants.

False (causal) relationship between facts (22%)

`Q` If there's an equal and opposite reaction for everything, how does any action happen? Isn't it balanced out by the opposite reaction? `C` I don't think you are fully comprehending what 'equal' means in this situation. (...) These forces are acting on different bodies so they do not cancel each other out. (...) `FP` The equal and opposite reaction applies to the same object.
`Q` In today's high tech world, how come we are not able to reproduce ancient crafting methods like Roman Concrete, Damascus Steel, or Greek Fire? `C` It's not that we can't reproduce them technologically, it's that the exact method or recipe was lost to history (...) `FP` Ancient crafting methods are not reproducible due to lack of technologies.

Exceptions (4%)

`Q` How do bugs and other insects survive winter when they have such a short lifespan? `C` Depends on the insect, some don't have that short of a lifespan. But mostly (...) `FP` (All) insects have a short lifespan.

No false presupposition / Annotation error (2%)

Table 2: Breakdown of types of false presuppositions, based on 50 random samples on the development data. `Q`, `C` and `FP` indicate the question, the comment, and the presupposition, respectively.

major confusion here", "I don't think you are fully comprehending ..." and "It should be noted that...". 80% of triggers appear in the first sentence of the comment, and the rest of the sentences elaborate on how the presupposition is false or provide other relevant information that does not directly answer the question. The rest 32% do not include lexical triggers and requires more careful comprehension of the comment with respect to the question, e.g., the false existential example in Table 2.

Analysis of ambiguous cases. Even with our best efforts, there are still inherent disagreement between annotators. Some of them are due to inherent ambiguities in language, e.g., the first example in Table 9 where 'the state of the water' could either mean the molecule itself or the energy state of the molecule. Others are due to disagreement on the validity of the presupposition, e.g., in the second example in Table 9, it is debatable whether or not the question writer presupposes that the Board deals with day to day at a company. We revisit this issue in human performance estimation in Section 5.2.

4 Task Setup

The model is given a question q, and is required to perform the following subtasks:

(a) Detection: assign a label to be `FP` or `N`; `FP` means q has a false presupposition, and `N` means q has no false presuppositions. We use a macro-F1 score as an evaluation metric.

(b) Writing: if `FP`, write the false presupposition as well as the correction. We use sacreBLEU (Post, 2018) and unigram-F1 following Petroni et al. (2021) as well as SentBERT (Reimers and Gurevych, 2019) for evaluation. We also introduce a human evaluation scheme in Section 6.3.

We have two tracks: the main track and the GOLD-COMMENT track.

The main track provides q as the only input to the model. The model is expected to search necessary background knowledge to perform the task from any information source except for Reddit and Quora.[8] This is the most realistic setting for the typical open-domain question answering problem.

The GOLD-COMMENT track provides the comment used for the annotation as an additional input to the model. This removes the need for retrieval, and guarantees that all necessary information to perform the task is given.

---

[8] This is because similar questions are being asked on these websites. The same decision has made in Nakano et al. (2021).

5 Experiments: Detection

This section discusses baseline experiments for the detection subtask; Section 6 discusses baseline experiments for the writing subtask.

5.1 Baselines

5.1.1 Trivial baselines

Random assigns `FP` or `N` randomly at uniform. **`FP` only** always assigns `FP`. **`N` only** always assigns `N`. **Nearest Neighbor** retrieves one of questions from the training set that is closest to the test question, based on c-REALM (Krishna et al., 2021), and returns its label as the prediction.

5.1.2 GOLD-COMMENT track baselines

Question only trains a RoBERTa-based (Liu et al., 2019) classifier that takes the question as the only input and classifies the label. It is often called closed-book model (Roberts et al., 2020). **Comment only** is a classifier based on RoBERTa-large that takes the comment as the only input and assigns the label. **Question⊕Comment** is a classifier based on RoBERTa-large that takes a concatenation of the question and the comment to the classifier, and assigns the label. We additionally experiment with the same model that is trained on either MNLI (Williams et al., 2018) or BoolQ (Clark et al., 2019), and tested on CREPE in a zero-shot fashion. This condition tests if training on similar, previously studied datasets helps.

5.1.3 Main track baselines

We design a model called **c-REALM + MP (Multi-passage) classifier** that retrieves a set of paragraphs from Wikipedia and then assigns a label.

First, the model uses c-REALM (Krishna et al., 2021), a state-of-the-art retrieval model on ELI5, to retrieve a set of k passages from the English Wikipedia. Next, the model uses the multi-passage classifier based on RoBERTa in order to assign a label. Given a question q and a set of passages $p_1 ... p_k$, each p_i $(1 \leq i \leq k)$ is concatenated with q and is transformed into $\mathbf{h}_i \in \mathbb{R}^h$ through the Transformer model. We then obtain logits via $\mathbf{p} = \mathrm{FFN}(\mathrm{MaxPool}(\mathbf{h}_1 ... \mathbf{h}_k)) \in \mathbb{R}^2$, where FFN is a feed-forward layer and $\mathrm{MaxPool}$ is an element-wise max operator. Finally, we use $\mathrm{Softmax}(\mathbf{p})$ to compute the likelihood of q having false presuppositions or not.

Self-labeling. Although our labeled data is small, there is large-scale unlabeled data (question and comment pairs) available. We explore self-labeling to leverage this unlabeled data. Specifically, we use the **Question⊕Comment** to assign a silver label to the unlabeled training questions. We then train the classifier on the union of this silver data as well as the gold labeled data.

5.1.4 Human performance

We estimate human performance to better understand the model performance. We recruit two human workers who perform the task for 186 questions for each track.

We estimate two types of human performance. (1) **Human with the most voted comment**, where human workers assume the most voted comment as a ground truth in terms of factuality of the information and the validity of the presupposition. We think of it as an upperbound of model performance. (2) **Human w/o the most voted comment**, where human workers search over the web (except Quora and Reddit) to find information, and make the best judgment about the validity of the presupposition. We think it is likely to be worse than the upperbound of model performance, since only one worker, instead of multiple online users or domain experts, makes a decision.

5.2 Results

Results are reported in Table 3.

The GOLD-COMMENT track. All trivial baselines achieve poor performance. In particular, poor performance of the nearest neighbor model indicates that there is no significant train-test overlap on CREPE. Using both the question and the comment (Question⊕Comment) achieves the best performance, outperforming the best trivial baseline by 22% absolute. Zero-shot models trained on MNLI and BoolQ achieve poor performance, indicating that our problem is significantly different from existing tasks like NLI or binary question answering. The best model is 10% below human performance, indicating room for improvement, even in the easier track.

The main track. Using retrieved passages from c-REALM and multi-passage classifier achieves 66.3% on the test set, which is significantly better than all trivial baselines. The self-labeling technique leads to additional improvements, leading to an F1 of 67.1%. While these numbers are significantly better than trivial baselines, they are significantly worse than the model performance given

Model	Dev	Test
Trivial baselines		
Random$^{\otimes}$	44.9	47.8
Always predict `FP`$^{\otimes}$	21.4	20.0
Always predict `N`$^{\otimes}$	42.1	42.9
Nearest Neighbor$^{\otimes}$	56.2	54.1
GOLD-COMMENT ***track***		
Question only	67.7	66.9
Comment only	68.9	68.6
Question$\oplus$Comment	**76.3**	**75.6**
Question$\oplus$Comment (MNLI)$^{\otimes}$	54.4	54.2
Question$\oplus$Comment (BoolQ)$^{\otimes}$	60.4	58.2
Main track		
c-REALM + MP classifier	68.3	66.3
c-REALM + MP classifier (Self-labeling)‡	**69.1**	**67.1**
Human w/ most-voted comment	86.4	85.1
Human w/o most-voted comment	70.9	70.9

Table 3: Baseline results in the **detection subtask** on the development data and the test data, respectively. **Macro-F1** scores reported. By default, the models are trained on the labeled portion of CREPE; $^{\otimes}$ indicates the model is not trained on CREPE; ‡ indicates the model is trained on both the labeled and unlabeled portions of CREPE.

the comment. This strongly suggests a retrieval bottleneck—getting passages that provide evidence as strong as human-written comments is difficult even with the state-of-the-art retrieval model.

To further support the bottleneck in retrieval, we conduct a detailed error analysis in Appendix C.2. For instance, 86% of false negatives were due to retrieval misses, including failing to retrieve relevant topics (42%), retrieving evidence on the relevant topics but not related to the false presuppositions (32%), or retrieving evidence related to the presuppositions but is not direct enough (12%).[9]

Human performance. Humans given the most upvoted comment achieve performance that is significantly higher than all baseline numbers, indicating significant room for improvement.

Without the most upvoted comment, people achieve relatively poor performance (70.9%). To better understand this, we analyze 44 error cases, and categorize them in Table 4. Nearly half of the errors are due to an inherent disagreement in labels, either due to (1) ambiguity, either in language or whether the presupposition was made, or (2) whether it is critical to correct false presuppositions (especially cases in the exception category in Table 2). We think using the most upvoted comment for a decision is reasonable since it is an aggregation of active community users and domain experts, but future work may take other approaches to consider ambiguities of the decision.

Error Category	%
Failure in finding evidence	11.4
Mistakes in labeling	11.4
Wrong ground truth label	11.4
Inherent disagreement: ambiguity	34.1
Inherent disagreement: criticalness	9.1
Information on the web being inconsistent	22.7

Table 4: Analysis of 44 errors made by human performers without the most upvoted comment.

6 Experiments: Writing

6.1 Baselines

In the writing subtask, the system is given a question that is guaranteed to contain a false presupposition, and is required to generate the presupposition as well as the correction.

6.1.1 GOLD-COMMENT track baselines

Copy baseline. As a trivial baseline, we copy the given question as a presupposition and the given comment as a correction.

Question$\oplus$Comment Dedicated. We train two generators separately to generate the presupposition and the correction, respectively, given a concatenation of the question and the comment. Both models are based on the pretrained T5-base model (Raffel et al., 2020).

Question$\oplus$Comment Unified. We also design a unified model that can be used for both the presupposition and the correction, motivated by the intuition that generation of each can benefit from each other. We train one generator that is trained with a union of (1) annotated corrections, and (2) annotated presuppositions prepended with "It is not the case that" so that they look like corrections. At inference time, we use a standard, beam search decoding to generate the correction. To generate the presupposition, we first decode a sequence with a constraint (De Cao et al., 2021) that it should start with "It is not the case that", and then take the sequence that comes next as a presupposition.

6.1.2 Main track baselines

We design **c-REALM + MP (Multi-Passage) Dedicated** and **c-REALM + MP (Multi-Passage) Unified**. They are similar to the dedicated and unified

---

[9] This can be seen as a classification error—if the classification model can better capture implicit evidence, it could have made a correct prediction.

Model	Development						Test					
	uF1			BLEU			uF1			BLEU		
	P	C	A	P	C	A	P	C	A	P	C	A
GOLD-COMMENT ***track***: Copy Baseline												
	45.0	26.5	35.8	14.8	5.5	10.2	44.7	27.0	35.9	14.6	5.7	10.2
GOLD-COMMENT ***track***: Question⊕Comment												
Dedicated	51.7	**37.8**	**44.8**	30.0	**17.9**	**24.0**	47.7	**38.0**	**42.9**	22.9	**16.1**	**19.5**
Unified	**53.4**	33.2	43.3	**30.9**	12.0	21.5	**49.2**	31.4	40.3	**25.9**	10.4	18.2
Main track: Question + c-REALM												
Dedicated	49.1	26.6	37.9	28.2	8.7	18.5	45.9	24.9	35.4	22.6	7.1	14.9
Unified	49.4	30.7	40.1	28.8	10.1	19.5	46.3	28.4	37.4	23.6	8.3	16.0

Table 5: Results in the **writing subtask** with **unigram F1 (uF1)** and **BLEU**. P, C, and A indicate Presupposition, Correction, and Average between the two.

Model	Development			Test		
	P	C	A	P	C	A
GOLD-COMMENT ***track***: Copy Baseline						
	75.5	70.8	73.2	75.6	70.9	73.3
GOLD-COMMENT ***track***: Question⊕Comment						
Dedicated	78.7	**75.7**	77.2	77.0	**75.2**	76.1
Unified	**82.7**	**75.7**	**79.2**	**82.0**	74.2	**78.1**
Main track: Question + c-REALM						
Dedicated	76.9	70.3	73.6	76.5	65.6	71.1
Unified	77.8	70.3	74.1	77.0	68.8	72.9

Table 6: Results in the **writing subtask** with **SentBERT**. P, C, and A indicate Presupposition, Correction, and Average between the two.

Model	F	P	CR	CS
GOLD-COMMENT ***track***: Question + Comment				
Dedicated	2.9	1.8	1.9	1.6
Unified	3.0	2.0	0.8	2.8
Main Track: Question + c-REALM				
Dedicated	2.8	1.8	0.6	1.6
Unified	3.0	1.8	0.6	2.8
Groundtruth	2.9	2.8	2.7	2.9
Agreement (%)	96.3	65.4	63.0	74.8

Table 7: Human evaluation results (scale: 0–3). F: Fluency, P: Presupposition, CR: Correction, CS: Consistency. The last row reports the inter-annotator agreement rate.

models in Section 6.1.1. The only difference is that the model receives a question and a set of k passages from c-REALM instead of a question-comment pair. In order for the T5 model to read multiple passages, we use the Fusion-in-Decoder architecture (Izacard and Grave, 2021). We refer to Appendix C.1 for more details.

6.2 Results: Automatic Evaluation

Table 5 reports the results in unigram F1 and BLEU. Examples of model outputs are provided in Appendix C.4. All models outperform the trivial copy baseline and perform better in the GOLD-COMMENT track than in the main track. Models are overall better at writing presuppositions than writing corrections, and the performance gap between the GOLD-COMMENT track and the main track is larger in presuppositions than in corrections. This is likely because the impact of evidence passages is more significant in correction writing than in presupposition writing since the presupposition can often be extracted from the question alone, while the correction requires information beyond the question. It is also worth noting that the unified model is better than the dedicated model in the main track but not in the GOLD-COMMENT track. This indicates that, while multi-task learning improves the main track, it does not improve the GOLD-COMMENT track, possibly because performing two tasks in isolation is sufficient.

Table 6 reports results in SentBERT scores. The overall scores are high, likely because SentBERT considers entailment between a reference and a generation. For instance, the copy baseline achieves high scores since the question and the presupposition, and the comment and the correction entail each other by definition. It is important to note that entailment is a necessary but not a sufficient condition for a presupposition or a correction to be satisfactory in our task definition. The next section shows that models with high SentBERT scores obtain low human ratings.

6.3 Results: Human Evaluation

We conduct human evaluation of model generations on 200 randomly sampled test instances from the following aspects (each in the 0–3 scale):

- **Fluency**: The generated text should be fluent (i.e., free of grammatical errors, spelling errors, and repetitions).
- **Presupposition**: The generated presupposition should be the valid one in the question, and is factually false.
- **Correction**: The correction should be made in the comment and provide reasonable amount of justification rather than being a simple negation of the presupposition.
- **Consistency**: The presupposition and correction should be on the same topic and negate each other.

We evaluate the output from all systems except the copying baseline, as well as the ground truth reference. Each question is assigned two raters in order to reduce noise and report inter-rater agreement on pairwise comparison. More details about the rating scheme are provided in Appendix C.4.

Based on results reported in Table 7, all models generate almost flawless fluent text and valid presuppositions. However, their outputs generated as false presuppositions are factually correct in half of the cases. These observations are relatively consistent across different systems.

Notable differences between systems are found in correction and consistency. The dedicated model generates better correction, likely because it is given a comment. All other models struggle: in particular, the unified models tend to generate the correction that starts with "It is not the case that" even the model is not restricted to do so at inference time. On the other hand, the unified model is better in consistency, likely because the dedicated model is more vulnerable in generating the presupposition and the correction in a totally different topic.

7 Conclusion

We introduced CREPE: the first benchmark for the identification and correction of false presuppositions in the open-domain setup. CREPE consists of 8,400 user questions, 25% of which contain false presuppositions and are paired with their corrections. Our detailed analysis highlights challenges in solving the task, including (1) retrieval of evidence that identifies false presupposition, (2) identification of implicit and subtle presuppositions, and (3) generating correction that is accurate and adequately explains how and why the presupposition is false. We hope our benchmark adds to the problem of open-domain, open-ended question answering, inviting researchers to build models to study questions with false presuppositions. Further, we suggest future work to develop better models, explore approaches to address inherent debatability of the judgment, and evaluation of the model generation.

Limitations

Inherent debatability in false presuppositions. As discussed earlier, the validity of presupposition is inherently debatable and largely depends on the background context, i.e., even experts in formal semantics and pragmatics observe a high disagreement rate (Jeretic et al., 2020). Our proposal in using the most upvoted comments partially address the issue, but not perfectly, as discussed extensively in Section 5.2. One avenue for future work is to consider extra-linguistic context such as individuals background when judging the validity of presuppositions (Zhang and Choi, 2021).

Evaluating massive language models. Massive language models such as GPT-3 (Brown et al., 2020) have been shown impressive performance in open-ended question answering (Nakano et al., 2021). Our paper does not include large-scale, systematic evaluation of such models. Instead, we conduct a small-scale case study with GPT-3 `text-davinci-002`. See Appendix D for details. Most generations are roughly on the right topic, but often contain information that is factually false and do not precisely answer the question. Moreover, they rarely explicitly identify false presupposition and provide corrections, indicating that GPT-3 is far from solving our task. We think future work may explore larger-scale evaluation in a more systematic manner.

False presuppositions beyond online forums. The domain of CREPE is limited to online forums (Reddit). While this choice was made due to the availability of large data and its general domain, we argue that false presuppositions are not specific to such domains. For instance, we find that a similar portion (25%) have false presuppositions on information-seeking questions on NLP research papers posed by NLP experts; see Appendix E for details. We think future work can explore creating benchmarks on such domains, as well as studying false presuppositions on a broader set of domains that require domain expertise.

Acknowledgements

We thank Daniel Fried and Julian Michael for their feedback on an early stage of the project. We thank Zeqiu Wu and H2Lab members for comments on the paper. We thank Li Du for suggesting the name of the dataset. This research is supported by DARPA MCS program through NIWC Pacific (N66001-19-2- 4031), NSF IIS-2044660, an Allen Distinguished Award and gifts from AI2. SM is supported by a J.P. Morgan fellowship.

References

Akari Asai and Eunsol Choi. 2021. Challenges in information-seeking QA: Unanswerable questions and paragraph retrieval. In *Proceedings of the Association for Computational Linguistics*.

FairScale authors. 2021. Fairscale: A general purpose modular pytorch library for high performance and large scale training. `https://github.com/facebookresearch/fairscale`.

David I. Beaver, Bart Geurts, and Kristie Denlinger. 2014. Presupposition. *The Stanford Encyclopedia of Philosophy*.

Rahul Bhagat and Eduard Hovy. 2013. Squibs: What is a paraphrase? *Computational Linguistics*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of Advances in Neural Information Processing Systems*.

Isabel Cachola, Eric Holgate, Daniel Preoţiuc-Pietro, and Junyi Jessy Li. 2018. Expressively vulgar: The socio-dynamics of vulgarity and its effects on sentiment analysis in social media. In *Proceedings of International Conference on Computational Linguistics*.

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

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of Empirical Methods in Natural Language Processing*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval. In *Proceedings of the International Conference on Learning Representations*.

Marie Duží and Martina Číhalová. 2015. Questions, answers, and presuppositions. *Computación y Sistemas*.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the Association for Computational Linguistics*. Association for Computational Linguistics.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The paraphrase database. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the European Chapter of the Association for Computational Linguistics*.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the Association for Computational Linguistics*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A Dataset of Peer Reviews (PeerRead): Collection, Insights and NLP Applications. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*.

S. Jerrold Kaplan. 1978. Indirect responses to loaded questions. In *Theoretical Issues in Natural Language Processing-2*.

Najoung Kim, Ellie Pavlick, Burcu Karagol Ayan, and Deepak Ramachandran. 2021. Which linguist invented the lightbulb? presupposition verification for question-answering. In *Proceedings of the Association for Computational Linguistics*.

Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to progress in long-form question answering. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*.

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*.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the Association for Computational Linguistics*. Association for Computational Linguistics.

Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2021. Question and answer test-train overlap in open-domain question answering datasets. In *Proceedings of the European Chapter of the Association for Computational Linguistics*.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint*.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. 2018. Mixed precision training. In *Proceedings of the International Conference on Learning Representations*.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of Empirical Methods in Natural Language Processing*.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint*.

Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2015. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the Association for Computational Linguistics*.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation*.

Christopher Potts. 2015. Presupposition and implicature. *The handbook of contemporary semantic theory*.

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*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the Association for Computational Linguistics*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of Empirical Methods in Natural Language Processing*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of Empirical Methods in Natural Language Processing*.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of Empirical Methods in Natural Language Processing*.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the Association for Computational Linguistics*.

John Schulman, Barret Zoph, Christina Kim, Jacob Hilton, Jacob Menick, Jiayi Weng, Juan Uribe, Liam Fedus, Luke Metz, et al. 2022. ChatGPT: Optimizing language models for dialogue.

Robert Stalnaker, Milton K Munitz, and Peter Unger. 1977. Pragmatic presuppositions. In *Proceedings of the Texas conference on per˜formatives, presuppositions, and implicatures.*

Peter F Strawson. 1950. On referring. *Mind*.

Ellen M. Voorhees and Dawn M. Tice. 2000. The TREC-8 question answering track. In *Proceedings of the Language Resources and Evaluation Conference*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*.

Fangyuan Xu, Junyi Jessy Li, and Eunsol Choi. 2022. How do we answer complex questions: Discourse structure of long-form answers. In *Proceedings of the Association for Computational Linguistics*.

Michael Zhang and Eunsol Choi. 2021. SituatedQA: Incorporating extra-linguistic contexts into QA. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7371–7387, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Details in Data Source

License The ELI5 (Fan et al., 2019) dataset uses the BSD License. Our usage of this dataset is consistent with its intended use, and the license will be included when downloading the data.

Filtering Data Sources. When the question has several comments, we choose the comment with the highest upvotes. Since our data is derived from Reddit, it may contain unintended social bias or harmful content. To remove toxic language on the data, we follow the toxicity word list from Cachola et al. (2018) to remove questions that contain any of the toxic words, except "hell" and "damn", as these two words are commonly used as interjections. The original authors of the ELI5 dataset (Fan et al., 2019) and we have made our best efforts to remove such content; however, it is possible that some harmful context may remain on the data after the filtering process.

Analysis in Train-Test Overlap. Krishna et al. (2021) reported that 81% of the validation questions of the original ELI5 dataset are paraphrases of the question on the training set. We revisit this issue, and show with a careful assessment of underlying assumptions and a finer-grained definition of "paraphrases", the proportion of paraphrased questions is significantly smaller.

We assess 104 randomly sampled questions in the validation set with their closest 7 training questions retrieved using c-REALM, as Krishna et al. (2021) did, but with a different rating scale (1–4), following Bhagat and Hovy (2013), Ganitkevitch et al. (2013), and Pavlick et al. (2015):

- **1: No Paraphrase; No similar intention**: Two questions do not share the meaning nor the intention.
- **2: No Paraphrase; Similar intention**: Two questions are likely to have the same intention, but they are not paraphrases, because their literal meanings are different and/or their underlying assumptions are different.
- **3: Non-trivial paraphrase:** Most of the two questions' meanings are the same; however, they do not belong to any of lexical paraphrase (single word to single word), phrasal paraphrase (multiword to single/multiword), or syntactic paraphrase (paraphrase rules containing non-terminal symbols),[10] and require non-trivial background knowledge to identify whether they have the same meaning.
- **4: Paraphrases**: Two questions fall into either lexical paraphrase, phrasal paraphrase, syntactic paraphrase, structured paraphrase, or other trivial paraphrase.

Table 8 presents the percentage and an example of each category. 76.9% questions have a rating of 2 or less, indicating that for most validation question, there are either no similar intention question in the training set, or there are questions with similar intention but either their literal meanings are different or their underlying assumptions are different. In particular, the latter indicates that whether or not there is a false presupposition can be different, even though they may share similar intention. Only 23.1% questions have a rating of 3 and above, indicating that relatively few questions in the validation set have a non-trivial or trivial paraphrase in the training set.

We also explored automatically filtering paraphrased questions using BLEU or TF-IDF. However, we find it is non-trivial to find the right threshold, thus include all questions and leave filtering to future work.

B Details in Data Annotation

The annotation instruction is in Figure 2, and we show an example of the annotation interface in Figure 3. The data collection is approved by an Institutional Review Board.

Qualification Task. The qualification task contains 20 pre-annotated questions by the authors, and we provide examples as well as their explanation to workers to demonstrate our task. Based on both whether the worker can correctly identify false presupposition and the writing quality, we selected 30 qualified workers.

Generation Task. Qualified workers who passed the qualification task work on the main annotation task. Each question is assigned to two generators who independently annotate the label. We monitor the agreement of workers and send the disagreed cases to further validate. We revoke qualification of generators whose more than 10% of annotations marked to be invalid by the validators.

Validation Task. If two generators disagree on the label, their annotations are sent to two validators who judge their validity. We select a smaller

---

[10] Definition derived from Ganitkevitch et al. (2013).

1 (45.2%)	***Dev***: Why are aluminium alloys difficult to weld? ***Train***: Cold welding. Two pieces of metal touch in a vacuum, why do they stick together? How strong is this weld?
2 (31.7%)	***Dev***: How is blood after a transfusion integrated into the body, especially when the transfused RBCs carry different DNA than the host RBCs? ***Train***: How DNA from blood is changed when getting a blood transfusion `comment` The dev question assumes that the transfused RBCs carry DNA, while the train question does not.
3 (6.7%)	***Dev***: How is information retained in solid-state memory devices after power is turned off? ***Train***: How do electronics keep memory after you take all power sources away? `comment` It is not trivial that "information retained in solid-state memory devices" is a paraphrase with "electronics keep memory".
4 (16.3%)	***Dev***: What is the difference between centrifugal and centripetal force? ***Train***: The difference between centrifugal force and centripetal force.

Table 8: The rating scale for paraphrase and examples for each category.

Ambiguity in language
`Q` When water boils, its bubbles are round, but when it freezes, its crystals are 6-sided. Why isn't frozen water round or boiling water hexagonally shaped? Aren't H2O molecules the same in either state? `C` Bubbles are round because (...) Ice crystals are shaped in such a way because (...) The water molecules are much slower and aren't bouncing all over the place. Gaseous H2O is much higher energy and further apart so that the regular pattern of ice doesn't come into effect.
Ambiguity in whether the presupposition is made
`Q` How do executives hold board of director positions at multiple fortune 500 companies? `C` The Board only meets occasionally to vote on more important matters. They don't really deal with day to day affairs like the CEO does.

Table 9: Two types of ambiguous cases. `Q` and `C` indicate the question and the comment, respectively.

number of high qualified workers who have exceptional understanding of false presuppositions and are active users of Reddit. We find that for a small number of highly ambiguous cases, two validators have disagreement. In this case, we send the question to a third validator and take the majority vote.

Generators and validators are paid with reasonable hourly wage (13 USD/hour and 18 USD/hour, respectively).

C Details in Experiments

C.1 Model details

Retrieval model. We use the English Wikipedia from 08/01/2019 provided by Petroni et al. (2021). We obtain c-REALM embeddings of Wikipedia passages as well as the questions. We use FAISS (Johnson et al., 2019) to do approximate maximum inner product search to retrieve the top-5 passages for the query.

Classifier model. For all the models in the GOLD-COMMENT track track, we use a per GPU batch size of 2, and for all the models in the main track, we use the top-5 passages as the context and a per GPU batch size of 8. We train all our model with learning rate 10^{-5}, weight decay of 0, and maximum sequence length of 256. We train all our models for 5000 steps and a gradient accumulation step of 5. We run the classifier models with 5 different random seeds and report the best result.

All experiments were done with two Nvidia-RTX6000 GPUs (the detection subtask) or two Nvidia-A40 GPUs (the writing subtask). We use half-precision training (Micikevicius et al., 2018) when training for our detection task and gradient checkpointing from fairscale (authors, 2021), and choose the best checkpoint based on the performance on the validation set and report its test set performance. For the unified models for the writing subtask, we choose the best checkpoint that gives the best BLEU score for presupposition.

C.2 Error analysis on the detection subtask

We conduct error analysis of predictions from c-REALM + MP classifier in the detection subtask. The authors annotate 50 false positive instances and 50 false negative instances on the validation set, sampled uniformly at random, and categorize the cause of errors.

Results are reported in in Table 10. Overall, most errors are due to failure in retrieval—although the model often successfully retrieves passages that are on right topic, it fails to retrieve passages that contain enough information about whether the backgrounded presuppositions are true or false. This issue is more prominent in false negatives, likely because finding that the presupposition is true requires significantly more exhaustive search than

Error Type		FP(%)	FN(%)
Retrieval	No related topic	26	32
	Similar topic but not enough to make decision	20	40
	Indirect evidence and require reasoning	–	12
Classification	Direct evidence	34	10
Labeling	Ground truth label is wrong	8	8
Inherent Disagreement	Ambiguity	4	2
	Criticalness	2	0
Inconsistency	Information on the web being inconsistent	10	4

Table 10: The breakdown analysis of false positive/false negative in the validation set for c-REALM + MP classifier model. FP: False positive. FN: False negative. `FP` : False Presupposition. "Indirect evidence and require reasoning" can also belong to retrieval error. Information on the web being inconsistent: the comment contradicts with the retrieved passages. The categories are *not* mutually exclusive.

`Q` Why aren't helicopters used to rescue hikers and remove bodies from Mt. Everest?
`C` The air in too thin and they can't fly that high up. They create lift by pushing air downwards. The higher up you go, the less air pressure you have, the less downward force a helicopter can make to fight gravity.
Retrieved Passage: In 2016 the increased use of helicopters was noted for increased efficiency and for hauling material over the deadly Khumbu icefall. In particular it was noted that flights saved icefall porters 80 trips but still increased commercial activity at Everest...
Original Label: No `FP`.
Model Prediction: `FP`.
`comment` Helicopter is used to rescue people from Mt. Everest according to the passages, but the comment does not point this `FP` out.

`Q` How does alcohol effect blood sugar in people with Diabetes?
`C` ...**Short answer: straight alcohol doesn't affect me at all (vodka, whiskey, etc).** I can drink without any noticeable effect on my blood sugar levels (and I have a continuous glucose monitor so I can literally see the effect or lack thereof)...
Retrieved Passage:The intake of alcohol causes an initial surge in blood sugar, and later tends to cause levels to fall. Also, certain drugs can increase or decrease glucose levels...
Original Label: `FP`.
Model Prediction: no `FP`.
`comment` The comment says that straight alcohol does not affect blood sugar change, but the retrieved passage says otherwise.

Table 11: Inconsistencies between comment and retrieved passages by c-REALM from Wikipedia.

noticing that the presupposition is false (since in most cases, whether the presupposition is correct is not explicitly mentioned).

Secondly, there are 8% cases of labeling error in false positive cases. Note that for false positive cases, we did not distinguish between direct evidence and indirect evidence because there is no clear definition of "evidence" for non `FP` cases. For false negative cases, if there is a retrieved passage that can directly *answer* the question but the model fails in labeling, we consider this as a classification error rather than retrieving error, because the goal for retrieval is to retrieve passages that could give the answer to the question. The model also suffers more to classify given indirect evidence that requires reasoning, which matches our intuition.

Inherent disagreement and inconsistency issues contribute to the rest of the errors. We found that minor part (4%, 2% for false positive and false negative cases, respectively) of the error due to the inherent disagreements in labeling ambiguity. We also found that 2% of the false positive errors are due to whether the `FP` is critical to satisfy users information need based on the question. Furthermore, we also found that 14% of the errors are due to the comment and the retrieved passages are not inconsistent, causing label mismatch, however, they are not a labeling error because our ground truth annotator is not presented with additional passages when annotating. Example of these inconsistencies can be found in Table 11.

C.3 Discussion on inherent ambiguity and inconsistency on the web

Table 12 display examples for the category "Inherent disagreement: ambiguity", "Inherent disagreement: criticalness", and "Information on the web being inconsistent" from the human evaluation section of Section 5.2 and in Table 4.

9.1% of the errors is due to the inherent ambiguity of the criticalness of the `FP`, i.e., whether correcting the `FP` is critical to satisfy users' information needed. For example, for the question "Why do things look darker when wet?" in Table 12, although our human rater found evidence on the internet that there exist things that look darker when dry, which would contradict the presupposition that (all) things look darker when wet, we believe that the question writer is mainly seeking an answer for the reason of the phenomenon that something looks darker when they are wet, and therefore, such `FP`is not critical to answer the users

Inherent disagreement: Ambiguity	**Information on the web being inconsistent**
`Q` How are paintings or museum art classified as masterpieces? Some look like paint scratches and yet they are classics. Why is that? `C` Great art isn't just technical skill. Though a lot of those paintings that may look easy usually aren't. But it's also about having an original idea and expressing yourself in a way the world hasn't seen before. (...) `comment` Whether the question presupposes that a painting or an art is considered a masterpiece only depends on the technical skill exist in the question is debatable without the comment. `Q` Why is Hydrogen so common on Earth and Helium quite rare? `C` Hydrogen is highly reactive, it bonds to oxygen, forming water. Water is quite dense, even as a vapor, and is therefore quite durable in the atmosphere. Helium is a noble gas and nearly perfectly inert. Being unbound to any heavier elements, it quickly rises to the top of the atmosphere and is lost to space by various mechanisms. Hydrogen is lost over time, but only slowly. `comment` There is an `FP` if the question is asking about hydrogen gas (see this webpage); however, if they are asking about hydrogen atom, there is no `FP`.	`Q` Why do 78% of Americans live paycheque to paycheck? ***News on 06/07/2022***: 58% Americans are living paycheck to paycheck. ***News on 01/11/2019***: 78% Americans are living paycheck to paycheck. `Q` Why do pandas eat bamboo and when did they stop eating meat? `C` Everyone so far has gone with the "pandas are so dumb" response so let me give you a proper answer. For a start, evolution is not a intelligent or forward thinking process. Bamboo may be low in energy, but it is abundant, grows quickly and not many other animals eat it. So for any animal that can evolve to consume it, there's an open niche there to be taken. Now I'll admit pandas aren't the best creature at surviving, but they don't really need to be. They live in an environment with abundant food and no predators or competitors, so all they need to do is sit around eating bamboo, expend minimal energy and produce just enough babies to keep the species going. Now that might not seem very efficient from a human perspective, but actually it's a strategy that works surprisingly well, and pandas were doing absolutely fine until humans came along and started hunting them and destroying their habitat. **WWF:** But they do branch out, with about 1% of their diet comprising other plants and even meat. While they are almost entirely vegetarian, pandas will sometimes hunt for pikas and other small rodents. **Science:** Pandas are one of the world's most fascinating vegetarians. Their digestive systems evolved to process meat, yet they eat nothing but bamboo—all day, every day. A new study reveals how these animals survive on a diet that should kill them.
Inherent disagreement: Criticalness	
`Q` Why do things look darker when wet? `C` You know how when you vacuum the floor it makes those different colored lines? If you look closely the darker colored parts of the carpet are laying down and the lighter colored parts are standing up. Many things that are dry have little hairs or rough surfaces that are basically standing up like little mirrors to reflect light. When these get knocked over or filled with water they can't reflect as well. A couple of damp riddles 1. What gets wetter as it dries 2. What gets darker as it dries	

Table 12: Inconsistencies and inherent ambiguity examples for human performance.

original question, and therefore the comment writer does not point it out.

Note that this is different than the "exception" examples mentioned in Table 2, as the comment explicitly pointed out the falsehood of the presupposition, and therefore we consider the `FP` as critical to answer the user's information seeking need.

Furthermore, 22.7% of the errors in human performance is due to information on the web being inconsistent. For the question "why do 78% of Americans live paycheque to paycheck?", News on 06/07/2022 points out that 58% Americans lives paycheck to paycheck, while News on 01/11/2019 pointed out that 78% of Americans live paycheck to paycheck. For the question "why do pandas eat bamboo and when did they stop eating meat?", creditable sources such as World Wide Fund(WWF) and Science says differently about whether panda eat meat or not.

34.1% of the errors are due to ambiguity, as analyzed in Section 3.3.

C.4 Details of the writing subtask

Example generations from each system are shown in Table 13, 14, and 15. As we discuss in Section 6.2, the unified model strongly prefers to generate the correction with prefix "It is not the case that", or simply repeat or negate the presupposition, even though we do not restrict it to do so at inference time.

Details in Automatic Evaluation. We use sacreBLEU (Post, 2018) from the datasets library (Lhoest et al., 2021). We follow the preprocessing and evaluation script of unigram-F1 following Petroni et al. (2021), and the evaluation script of SentBERT following Reimers and Gurevych (2019).

Details in Human Evaluation. We conduct human evaluation for 200 questions on the test data. We recruit two student evaluators, who independently evaluate the presupposition and the correction given a question and its comment. They evalu-

`Inputs given to the human raters`

Question: Why do prosecuters/courts seek/sentence prison time greater than the expected lifespan of the offender (i.e. 150 years in prison)? Why not simply sentence those criminals to 'life' in prison instead?

Comment: Sentencing options are written into state laws. Life in prison is different in state laws than 150 years. Some of it comes into play with the "cruel and unusual punishment" clause in the Constitution too. Life in prison may not be "cruel and unusual" for a murder sentence, but it might be for, say, child sex trafficking. But if you trafficked 10 kids and the sentence is 15 years for each one, you get an effective life sentence that will also stand up, Constitutionally, against a "cruel and unusual punishment" defense.

`Outputs human raters rate`

Reference

Presupposition: It does not make sense to sentence a person to 150 years in prison if they can't live that long anyways, prosecutors should use the life in prison sentence instead.

Correction: The defendant can argue the life in prison sentence as cruel and unusual, so the actual year sentence is better to give than the alternative.

GOLD-COMMENT track, Dedicated

Presupposition: Penalties should be able to be sentenced to life in prison.

Correction: Life in prison is different in state laws than 150 years in prison.

GOLD-COMMENT track, Unified

Presupposition: If a criminal is sentenced to life in prison, they should be sentenced to life in prison.

Correction: It is not the case that if a criminal is sentenced to life in prison, they should be sentenced to life in prison.

Main, Dedicated

Presupposition: Penalties should be able to be imposed on criminals for life.

Correction: The longer the sentence, the more likely the prosecution will seek to sentence the offender to life in prison.

Main, Unified

Presupposition: Prosecutor's should seek prison time greater than the expected lifespan of the offender.

Correction: It is not the case that prosecutor's should seek prison time greater than the expected lifespan of the offender.

Table 13: An example of the input and the output human raters are given for the human evaluation of the writing subtask. Note that human raters are not given which output is a reference or from which system.

`Inputs given to the human raters`

Question: Why did scientists in the 1970s think that there was going to be a new ice age soon?

Comment: They didn't. Between 1965 and 1979, there was 7 papers talking about global cooling (not ice age and not necessarily soon). During the same period there was 44 papers about global warming. The media just liked the sensationalism, so there was some news article and a front page on the Times Magazine. They started with a minority of scientist talking about global cooling in a time period when there was still a lot of unknown in climate science and changed that to Scientific consensus that an Ice Age is coming soon. The 7 papers were the following : McComick and Ludwig 1967, Barrett 1971, Rasool and Schneider 1971, Hamilton and Seliga 1972, Chylek and Coakley 1974, Bryson and Dittbemer 1976 and Twomey 1977.

`Outputs human raters rate`

Reference

Presupposition: Scientists in the 1970s had a belief of an upcoming ice age.

Correction: The scientists of this time talked of global cooling in their near future, however they did not refer to this as an ice age.

GOLD-COMMENT track, Dedicated

Presupposition: Scientists thought there would be a new ice age soon.

Correction: Scientists didn't think that there would be a new ice age soon, they just thought that there would be one.

GOLD-COMMENT track, Unified

Presupposition: Scientists thought that there was going to be a new ice age soon.

Correction: Scientists didn't think there was going to be a new ice age soon.

Main, Dedicated

Presupposition: Scientists in the 1970s thought that there was going to be a new ice age soon.

Correction: Scientists in the 1970s did not think that there was going to be a new ice age.

Main, Unified

Presupposition: Scientists in the 1970s thought that there was going to be a new ice age soon.

Correction: Scientists in the 1970s thought that there was going to be a new ice age soon.

Table 14: Another example of the input and the output human raters are given for the human evaluation of the writing subtask. Note that human raters are not given which output is a reference or from which system.

ated five outputs, including the reference in the data as well as generations from four systems in Section 6.1: the dedicated and the unified model from the GOLD-COMMENT track and the main track, respectively.

We design detailed evaluation scheme, hold a 1 hour in-person tutorial session for human evaluators to be familiarized with the evaluation task.

`Inputs given to the human raters`

Question: If the phases of the moon are due to the Earths shadow, why is the half moon a straight line/perfect semi circle, instead of a rounded shadow?

Comment: Unless you're looking at a lunar eclipse, the part of the moon that's not lit up is not in the Earth's shadow, it's in the moon's shadow. The bright portion is facing the sun, and is therefore illuminated, and the dark portion is facing away from the sun, and therefore not illuminated.

`Outputs human raters rate`

Reference

Presupposition: If the phases of the moon are presumed to be the Earth's shadow, then phases should never be a straight line due to the earth being rounded.

Correction: With the exception of the lunar eclipse, the phases of the moon are not caused by the shadow of the Earth.

GOLD-COMMENT track, Dedicated

Presupposition: The phases of the moon are due to the Earth's shadow.

Correction: The part of the moon that is not lit up is in the moon's shadow.

GOLD-COMMENT track, Unified

Presupposition: The phases of the moon are due to the Earth's shadow.

Correction: It is not the case that the phases of the moon are due to the Earth's shadow.

Main, Dedicated

Presupposition: Half moon is a straight line/perfect semi circle.

Correction: The moon is not a rounded shadow, it is a straight line.

Main, Unified

Presupposition: The phases of the moon are due to the Earth's shadow.

Correction: It is not the case that the phases of the moon are due to the Earth's shadow.

Table 15: Another example of the input and the output human raters are given for the human evaluation of the writing subtask. Note that human raters are not given which output is a reference or from which system.

In particular, each output is rated based on four aspects as follows.

Fluency measures the fluency of the generated text, mainly whether it have repetitions, spelling errors or grammatical errors, or gibberish.

0: Generated text have fluency errors.
3: Generated text is free of fluency errors.

Presupposition evaluates whether the generated presupposition is the valid one in the question and whether it is factually false according to the comment.

0: The presupposition is invalid, i.e., does not exist in the question.
1: The presupposition is valid, e.g., exists in the question, but it is not factually false.
3: The presupposition is valid and is factually false.

Correction evaluates whether the generated correction provides a valid correction to the presupposition based on the comment with no hallucinated information, and provides enough justification (rather than simply being a negated presupposition). The former considers correctness (precision of the information), while the latter considers adequacy (recall of the information)

0: The correction is wrong based on the comment, or the correction is hallucinated.
1: The correction is correct based on the comment, but no additional information is provided to justify, or is a simple negation of the presupposition.
2: The correction is correct based on the comment, but misses some details to fully justify the falsehood of presupposition.
3: The correction is correct and provide enough information to justify the falsehood of presupposition.

Consistency requires the generated assumption and correction should be on the same topic and negate each other.

0: The presupposition and correction are not about the same topic.
1: The presupposition and correction are on the same topic, but they are not negating each other, or the negation is not explicit.
3: The presupposition and correction are consistent: are on the same topic and negate each other.

The evaluators are paid 17 USD/hour. See Section 6.3 for the results and discussion.

D A case study with GPT-3

Large language models such as GPT-3 (Brown et al., 2020) have shown impressive performance

Q: If there's an equal and opposite reaction for everything, how does any action happen? Isn't it balanced out by the opposite reaction?
FP: The equal and opposite reaction apply to the same object.
Corr: Based on Newton's Law of Motion, the equal and opposite reaction apply to different objects. Only forces that are applied to the same object would be cancelled out.

1 GPT-3 given Q, temp=0.0: The equal and opposite reaction is not always instantaneous, so an action can still happen even if there is an equal and opposite reaction.
`comment` Factually false. The equal and opposite reaction are always instantaneous.
2 GPT-3 given Q, temp=0.7: According to Newton's third law of motion, "For every action, there is an equal and opposite reaction." However, this only applies to objects interacting with other objects. In other words, if Object A exerts a force on Object B, then Object B will exert an equal and opposite force on Object A. However, if there is no other object for the force to be exerted on, then there is no equal and opposite reaction, and the force will simply cause the object to accelerate in the direction of the force.
`comment` Does not explicitly point out `FP`, and the correction of `FP` is factually wrong.
3 GPT-3 given Q, temp=0.7 (different sample): There is an equal and opposite reaction for every action, but this doesn't mean that every action is balanced out by the opposite reaction. For example, if you push on a wall, the wall pushes back on you with an equal and opposite force. However, the wall doesn't move because the force you're exerting on it is balanced out by the force of the wall pushing back on you.
`comment` The correction of `FP` is factually wrong. In the scenario mentioned in the generation, the wall does not move because its mass is large. Besides, this scenario is not related to what the question is asking about.
4 GPT-3 given Q + "Or did I miss anything?", temp=0.7: No, you didn't miss anything. The equal and opposite reactions do balance each other out, but the net result is still an action. This is because the reactions don't happen instantaneously - there is a brief period of time during which one reaction is occurring without the other. During this time, the system is unbalanced and an action can occur.
`comment` Factually false, and fail to identify `FP`. The equal and opposite reactions do not balance out, and they happen instantaneously.
5 GPT-3 given Q + "Or did I misunderstand anything?", temp=0.7: Newton's third law of motion states that for every action, there is an equal and opposite reaction. However, this does not mean that every action is perfectly balanced by an equal and opposite reaction. Instead, it means that the total force exerted by all objects on each other is always zero.
`comment` Does not explicitly point out `FP`, and provide adequate correction.
6 GPT-3 given Q + "Or do I have any misunderstanding?", temp=0.7: There is an equal and opposite reaction for every action, but not every reaction is equal in magnitude to the action. For example, if you push on a wall, the wall pushes back on you with an equal and opposite force, but the force exerted by the wall is much less than the force you exerted. This is because the wall is much more massive than you are, so it takes more force to move it.
`comment` Factually false. The force exerted by 'you' and the force exerted by the wall is equal. Besides, this is not what the question is asking about. It is answering a different question: why the wall won't move given a force from you.

Table 16: Response from InstructGPT (`text-davinci-002`) given an example question. temp=0.0 is deterministic, and temp=0.7 is the default value in GPT-3. Generations are obtained by sampling multiple times or using slightly different templates. We did not cherry-pick these generations: we ran the model six times and report all.

Q: How does flavored seltzer have 0 calories per serving but 10 calories per bottle?
FP: If something has 0 calories per serving then it cannot have any calories in total.
Corr: If a serving has under 5 calories, it can be rounded down to 0.

Instruct GPT-3: Flavored seltzer typically has 0 calories per serving, but 10 calories per bottle because the bottle contains multiple servings.
`comment` Does not point out that there can be > 0 calories per serving.
ChatGPT `Mar 08th Snapshot`: ... The reason that a bottle of flavored seltzer may list 10 calories on the label is because the servicing size is often one full bottle, which typically contains multiple servings. So, while a single serving may have 0 calories, the entire bottle may contain 10 calories or more....
`comment` Does not point out that there can be > 0 calories per serving.
ChatGPT `May 21th Snapshot`: ... The discrepancy you mentioned regarding flavored seltzer's calorie content per serving and per bottle is likely due to serving size and rounding conventions. When a flavored seltzer is labeled as having 0 calories per serving, it means that the amount of calories present in a single serving is so low that it falls below the threshold required for rounding up to the nearest whole calorie. This is typically 5 calories per serving. Consequently, the manufacturer can round down and label it as having 0 calories per serving...
`comment` Correct.

Table 17: Response from Instruct GPT-3 (`text-davinci-002`) and ChatGPT (Schulman et al. (2022), `March snapshot`, `May snapshot`) given an example question.

in generating a response to the question. We conduct small-scale evaluation of Instruct GPT-3 (`text-davinci-002`), whose details are not public but is known as the best version of GPT-3. An example is depicted in Table 16. We find that most generations are roughly on the right topic, e.g., all generations in Table 16 discuss Newton's Law of Motion. However, they rarely correctly satisfy users information need:

- Most of them include information that is factually false, e.g., the equal and the opposite action are not instantaneous (1 4), their magnitude is not equal (2 6), or they do balance out (3).
- They are often not precisely about what the question is asking about. For instance, they discuss why an object may not move given a force, e.g., the wall does not move when you hit the wall (2 6). This is related to Newton's Law of Motion, but not at all to the question.
- They do not explicitly identify false presuppositions. None of the generation mentions that the key misunderstanding is that the equal and opposite reaction apply to different objects, thus are not cancelled out. Sometimes the generation indicate some part of the question is wrong (indicates with 'but' or 'However') but does not precisely point out what is wrong, nor provide corrections.

It is possible performance could be improved with better prompting, but we leave this possibility to future work. We also experiment with ChatGPT in Mar 2023 and May 2023 in Table 17, but due to the closeness and the continuation of update, it is hard to evaluate and do an ablation study of the model, but we think the model is better than InstructGPT in generating answers that explicitly point out the false presupposition, but we leave the possibility to systematically evaluate this model on our task to future work.

E False Presuppositions in Other Data

While we focus on questions from an online forum due to the availability of large unlabeled data and the domain being fairly general, we argue that false presuppositions are not specific to such domains. In fact, false presuppositions are more prevalent when the domain is specific and requires expertise.

We analyze 50 random samples of unanswerable questions from QASPER (Dasigi et al., 2021), a dataset consisting of information-seeking questions on NLP research papers, posed by NLP experts. We find that, out of 48 questions that are unanswerable (2 of them turn out to have valid answers), 25% of them has false presuppositions, because the question writer does not have sufficient background knowledge or misses facts in the research paper. Table 18 shows a subset of such questions, along with false presuppositions and their correction annotated by the author. Identification of false presuppositions requires external knowledge beyond the specific paper the question writer is reading (1), or requires understanding of details of the paper which may not be explicitly written with the exact same terms in the paper (2 and 3).

---

1 ***Paper title:*** Combating Adversarial Misspellings with Robust Word Recognition
Q: Why do they experiment with RNNs instead of transformers for this task?
FP: The paper does not experiment with transformers.
Corr: The paper uses RNNs and BERT. The question writer either missed the fact that they used BERT, or did not know that BERT is based on transformers.

---

2 ***Paper title:*** Analysis of Wikipedia-based Corpora for Question Answering
Q: Can their indexing-based method be applied to create other QA datasets in other domains, and not just Wikipedia?
FP: Their indexing-based method is applied to create a QA dataset in the Wikipedia domain.
Corr: Their indexing-based method is not for creating QA datasets. This is for aligning (already annotated) answer context to a particular Wikipedia corpus.

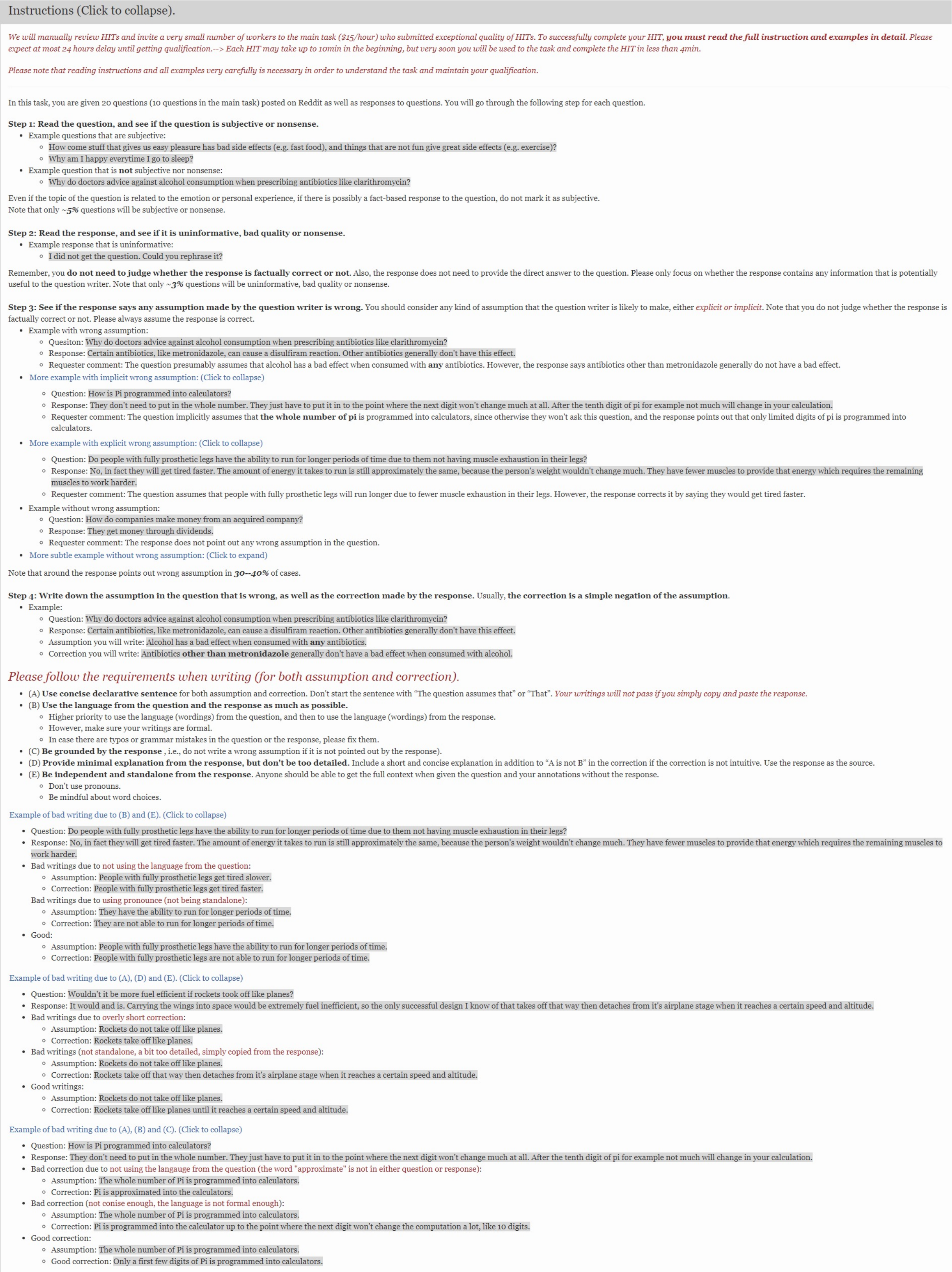
---

3 ***Paper title:*** Automatic Classification of Pathology Reports using TF-IDF Features
Q: How many annotators participated?
FP: There are annotators.
Corr: There is no annotators. The paper created a dataset, but the data construction process is entirely automatic.

Table 18: Example questions with false presuppositions on QASPER (Dasigi et al., 2021). *Q*, *FP* and *Corr* indicate the question, false presupposition, and the correction, respectively.

It should be noted that this percentage is a strict lower bound of true percentage and may be significantly underestimated since identification of false presupposition requires knowledge in NLP; the estimate will be higher when annotated by multiple NLP experts. Moreover, presuppositions will significantly more prevalent when question writers are non-expertise, unlike in QASPER whose question writers are NLP experts.

We did not annotate and experiment with QASPER because the data is relatively small (272 and 62 unanswerable questions on the training set and on the validation set, respectively), but future work can investigate this data as well we domain transfer of models.



Instructions (Click to collapse).

*We will manually review HITs and invite a very small number of workers to the main task ($15/hour) who submitted exceptional quality of HITs. To successfully complete your HIT, **you must read the full instruction and examples in detail**. Please expect at most 24 hours delay until getting qualification.--> Each HIT may take up to 10min in the beginning, but very soon you will be used to the task and complete the HIT in less than 4min.*

Please note that reading instructions and all examples very carefully is necessary in order to understand the task and maintain your qualification.

In this task, you are given 20 questions (10 questions in the main task) posted on Reddit as well as responses to questions. You will go through the following step for each question.

Step 1: Read the question, and see if the question is subjective or nonsense.

- Example questions that are subjective:
  - How come stuff that gives us easy pleasure has bad side effects (e.g. fast food), and things that are not fun give great side effects (e.g. exercise)?
  - Why am I happy everytime I go to sleep?
- Example question that is **not** subjective nor nonsense:
  - Why do doctors advice against alcohol consumption when prescribing antibiotics like clarithromycin?

Even if the topic of the question is related to the emotion or personal experience, if there is possibly a fact-based response to the question, do not mark it as subjective.
Note that only ~***5%*** questions will be subjective or nonsense.

Step 2: Read the response, and see if it is uninformative, bad quality or nonsense.

- Example response that is uninformative:
  - I did not get the question. Could you rephrase it?

Remember, you **do not need to judge whether the response is factually correct or not**. Also, the response does not need to provide the direct answer to the question. Please only focus on whether the response contains any information that is potentially useful to the question writer. Note that only ~***3%*** questions will be uninformative, bad quality or nonsense.

Step 3: See if the response says any assumption made by the question writer is wrong. You should consider any kind of assumption that the question writer is likely to make, either *explicit or implicit*. Note that you do not judge whether the response is factually correct or not. Please always assume the response is correct.

- Example with wrong assumption:
  - Quesiton: Why do doctors advice against alcohol consumption when prescribing antibiotics like clarithromycin?
  - Response: Certain antibiotics, like metronidazole, can cause a disulfiram reaction. Other antibiotics generally don't have this effect.
  - Requester comment: The question presumably assumes that alcohol has a bad effect when consumed with **any** antibiotics. However, the response says antibiotics other than metronidazole generally do not have a bad effect.
- More example with implicit wrong assumption: (Click to collapse)
  - Question: How is Pi programmed into calculators?
  - Response: They don't need to put in the whole number. They just have to put it in to the point where the next digit won't change much at all. After the tenth digit of pi for example not much will change in your calculation.
  - Requester comment: The question implicitly assumes that **the whole number of pi** is programmed into calculators, since otherwise they won't ask this question, and the response points out that only limited digits of pi is programmed into calculators.
- More example with explicit wrong assumption: (Click to collapse)
  - Question: Do people with fully prosthetic legs have the ability to run for longer periods of time due to them not having muscle exhaustion in their legs?
  - Response: No, in fact they will get tired faster. The amount of energy it takes to run is still approximately the same, because the person's weight wouldn't change much. They have fewer muscles to provide that energy which requires the remaining muscles to work harder.
  - Requester comment: The question assumes that people with fully prosthetic legs will run longer due to fewer muscle exhaustion in their legs. However, the response corrects it by saying they would get tired faster.
- Example without wrong assumption:
  - Question: How do companies make money from an acquired company?
  - Response: They get money through dividends.
  - Requester comment: The response does not point out any wrong assumption in the question.
- More subtle example without wrong assumption: (Click to expand)

Note that around the response points out wrong assumption in ***30--40%*** of cases.

Step 4: Write down the assumption in the question that is wrong, as well as the correction made by the response. Usually, **the correction is a simple negation of the assumption**.

- Example:
  - Question: Why do doctors advice against alcohol consumption when prescribing antibiotics like clarithromycin?
  - Response: Certain antibiotics, like metronidazole, can cause a disulfiram reaction. Other antibiotics generally don't have this effect.
  - Assumption you will write: Alcohol has a bad effect when consumed with **any** antibiotics.
  - Correction you will write: Antibiotics **other than metronidazole** generally don't have a bad effect when consumed with alcohol.

Please follow the requirements when writing (for both assumption and correction).

- (A) **Use concise declarative sentence** for both assumption and correction. Don't start the sentence with "The question assumes that" or "That". *Your writings will not pass if you simply copy and paste the response.*
- (B) **Use the language from the question and the response as much as possible.**
  - Higher priority to use the language (wordings) from the question, and then to use the language (wordings) from the response.
  - However, make sure your writings are formal.
  - In case there are typos or grammar mistakes in the question or the response, please fix them.
- (C) **Be grounded by the response** , i.e., do not write a wrong assumption if it is not pointed out by the response).
- (D) **Provide minimal explanation from the response, but don't be too detailed.** Include a short and concise explanation in addition to "A is not B" in the correction if the correction is not intuitive. Use the response as the source.
- (E) **Be independent and standalone from the response**. Anyone should be able to get the full context when given the question and your annotations without the response.
  - Don't use pronouns.
  - Be mindful about word choices.

Example of bad writing due to (B) and (E). (Click to collapse)

- Question: Do people with fully prosthetic legs have the ability to run for longer periods of time due to them not having muscle exhaustion in their legs?
- Response: No, in fact they will get tired faster. The amount of energy it takes to run is still approximately the same, because the person's weight wouldn't change much. They have fewer muscles to provide that energy which requires the remaining muscles to work harder.
- Bad writings due to not using the language from the question:
  - Assumption: People with fully prosthetic legs get tired slower.
  - Correction: People with fully prosthetic legs get tired faster.

  Bad writings due to using pronounce (not being standalone):
  - Assumption: They have the ability to run for longer periods of time.
  - Correction: They are not able to run for longer periods of time.
- Good:
  - Assumption: People with fully prosthetic legs have the ability to run for longer periods of time.
  - Correction: People with fully prosthetic legs are not able to run for longer periods of time.

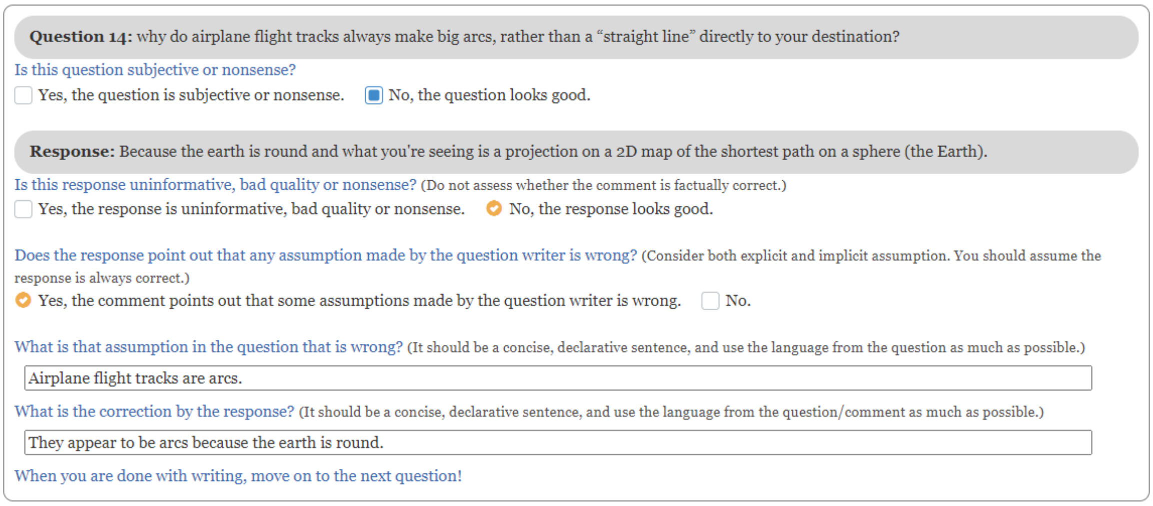
Example of bad writing due to (A), (D) and (E). (Click to collapse)

- Question: Wouldn't it be more fuel efficient if rockets took off like planes?
- Response: It would and is. Carrying the wings into space would be extremely fuel inefficient, so the only successful design I know of that takes off that way then detaches from it's airplane stage when it reaches a certain speed and altitude.
- Bad writings due to overly short correction:
  - Assumption: Rockets do not take off like planes.
  - Correction: Rockets take off like planes.
- Bad writings (not standalone, a bit too detailed, simply copied from the response):
  - Assumption: Rockets do not take off like planes.
  - Correction: Rockets take off that way then detaches from it's airplane stage when it reaches a certain speed and altitude.
- Good writings:
  - Assumption: Rockets do not take off like planes.
  - Correction: Rockets take off like planes until it reaches a certain speed and altitude.

Example of bad writing due to (A), (B) and (C). (Click to collapse)

- Question: How is Pi programmed into calculators?
- Response: They don't need to put in the whole number. They just have to put it in to the point where the next digit won't change much at all. After the tenth digit of pi for example not much will change in your calculation.
- Bad correction due to not using the langauge from the question (the word "approximate" is not in either question or response):
  - Assumption: The whole number of Pi is programmed into calculators.
  - Correction: Pi is approximated into the calculators.
- Bad correction (not conise enough, the language is not formal enough):
  - Assumption: The whole number of Pi is programmed into calculators.
  - Correction: Pi is programmed into the calculator up to the point where the next digit won't change the computation a lot, like 10 digits.
- Good correction:
  - Assumption: The whole number of Pi is programmed into calculators.
  - Good correction: Only a first few digits of Pi is programmed into calculators.

Figure 2: The instruction we provided for our qualification task.



Question 14: why do airplane flight tracks always make big arcs, rather than a "straight line" directly to your destination?

Is this question subjective or nonsense?

☐ Yes, the question is subjective or nonsense. ☑ No, the question looks good.

Response: Because the earth is round and what you're seeing is a projection on a 2D map of the shortest path on a sphere (the Earth).

Is this response uninformative, bad quality or nonsense? (Do not assess whether the comment is factually correct.)

☐ Yes, the response is uninformative, bad quality or nonsense. ☑ No, the response looks good.

Does the response point out that any assumption made by the question writer is wrong? (Consider both explicit and implicit assumption. You should assume the response is always correct.)

☑ Yes, the comment points out that some assumptions made by the question writer is wrong. ☐ No.

What is that assumption in the question that is wrong? (It should be a concise, declarative sentence, and use the language from the question as much as possible.)

Airplane flight tracks are arcs.

What is the correction by the response? (It should be a concise, declarative sentence, and use the language from the question/comment as much as possible.)

They appear to be arcs because the earth is round.

When you are done with writing, move on to the next question!

Figure 3: An example of our designed annotation interface.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
In limitations, Appendix D.2, D.3

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

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

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

B ☑ Did you use or create scientific artifacts?

Section 3

☑ B1. Did you cite the creators of artifacts you used?
Section 3.1, Data source

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Appendix A. We will add license in our released github repository so that people will download the data with the license.

☑ 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)?
Intend use of the data is clear from Section 1, 3, and 4.

☑ 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?
Appendix A: Details in Data Source includes how we remove toxic languages in the dataset. No personal information other than MTurk ID is collected in the data collection process.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Table 2, Section 1 L121-125.

☑ 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.
Table 1

C ☑ Did you run computational experiments?

Section 5 and 6.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 5.1, Section 5.2, Appendix D.1

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?
Appendix D.1

☑ 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?
Appendix D.1

☑ 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.)?
Section 4, Appendix D.4

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

Section 3, Appendix B

☑ 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.?
Appendix D.4., Figure 2, Figure 3

☑ 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)?
Appendix B, Appendix D.4

☒ 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?
We recruit people and people need to opt-in to participate in annotation. All our annotators were aware of the intent of the annotations.

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Appendix B

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Appendix D.4