Open Domain Question Answering with Conflicting Contexts

Anonymous ACL submission

Abstract

Open domain question answering systems frequently rely on information retrieved from large collections of text (such as the Web) to answer questions. However, such collections of text often contain conflicting information, and indiscriminately depending on this information may result in untruthful and inaccurate answers. To understand the gravity of this problem, we collect a human-annotated dataset, *Question Answering with Conflicting Contexts (QACC)*, and find that as much as 25% of *unambiguous*, open domain questions can lead to conflicting contexts when retrieved using Google Search. We evaluate and benchmark three powerful Large Language Models (LLMs) with our dataset *QACC* and demonstrate their limitations in effectively addressing questions with conflicting information. To explore how humans reason through conflicting contexts, we request our annotators to provide explanations for their selections of correct answers. We demonstrate that by finetuning LLMs to explain their answers, we can introduce richer information into their training that guide them through the process of reasoning with conflicting contexts. We will release our dataset and code to promote research along this line.

1 Introduction

Large language models (LLMs) have shown impressive capabilities on question answering tasks. In an open domain setting, a typical approach involves (1) retrieving relevant documents as contexts from the web or knowledge bases, and (2) using LLMs to generate the answer with the guide of the context. However, retrieved contexts from the web could often present **conflicting** information: *e.g.,* 22.62% pregnant women reported to find conflicting medical information from different websites in a survey(Hämeen-Anttila et al., 2014), such conflicts can lead to undesirable consequences when a language model relies indiscriminately on them to answer questions.

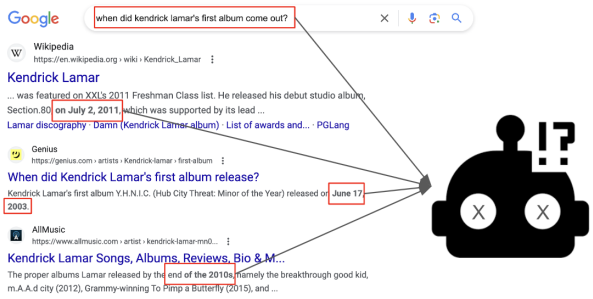


Figure 1: Google search results when querying the question *"When did Kendrick Lamars first album come out?"*. We can see that here different answers (July 2, 2011 / June 17, 2003 / end of the 2010s) are suggested by Google and it is difficult for a language model to decide which to believe in.

Previous work have explored different aspects of conflicts in the field of Natural Language Processing (NLP), including having different perspectives (Chen et al., 2019; Liu et al., 2021), fake news and misinformation (Chen et al., 2022b; Pan et al., 2023), conflicts due to ambiguous or inadequate questions (Min et al., 2020; Zhang and Choi, 2021), knowledge that change over time (Kasai et al., 2023), and knowledge clashes between the parameters and contexts (Longpre et al., 2021; Chen et al., 2022a; Xie et al., 2024).

In this work, we target the **conflicts among the contexts** when retrieving from the web with an *unambiguous* query and study their impact on the downstream question answering task. Figure 1 shows the results of querying *"when did kendrick lamars first album come out?"* on Google[1]. We observe that in the top-10 returned results, there are different evidence suggesting different answers to the question, and such inconsistencies may confuse the language models when they refer to the contexts to answer the question. Earlier works study this

[1]Results were queried in June 2024.

issue of conflicting contexts through perturbations with entity-substitution (Chen et al., 2022a; Hong et al., 2024), machine-generation (Pan et al., 2023; Wan et al., 2024; Hong et al., 2024), rule-based templates (Kazemi et al., 2023), or on controversial, multi-perspective questions (Liu et al., 2021; Wan et al., 2024). However, none of them examine the scenario where realistic, unambiguous open domain questions can also lead to conflicting contexts on the web and its effect in downstream question answering.

To quantify how often conflicting contexts occur on the web, we construct our dataset named QACC (**Q**uestion **A**nswering with **C**onflicting **C**ontexts). We consider *unambiguous* open domain questions from AmbigQA (Min et al., 2020) and use Google Search API[2] to retrieve up to 10 results for each question. We then use Amazon Mechanical Turk and ask human annotators to determine whether there exists different answers in the contexts. We find that about 25% of the *unambiguous* open domain questions will yield conflicting evidence from Google. We evaluate three popular LLMs (GPT-4o, Claude-3, and Phi-3) on our dataset with different prompting and finetuning strategies and establish that conflicting contexts can lead to substantial performance degradation in them. To understand how humans reason through conflicting contexts, we ask our annotators to select from a pre-defined set of reasons when deciding on the answer. Our findings indicate that humans often adhere to majority vote (i.e. selecting the most popular answer) when seeing conflicting contexts. In addition, we also request our annotators to provide a single sentence, natural language explanation for their answers. We find that by finetuning LLMs to explain their answers, we can introduce richer information into their training that guide them through the process of reasoning with conflicting contexts and improve their performance in both QACC and a perturbed NQ-Open dataset (Lee et al., 2019).

To summarize, our contributions in this work are the following:

- We construct a human-annotated dataset QACC and find that about 25% of unambiguous, open domain questions can lead to conflicting contexts when queried with Google Search.
- We benchmark open domain question answering with conflicting contexts with our dataset QACC and demonstrate the limitations of current LLMs under this scenario.
- We show that when finetuning with human explanations, LLMs can improve their abilities to answer questions correctly with conflicting contexts.

[2] https://developers.google.com/custom-search/v1/overview

2 Related Work

2.1 Retrieval Augmented Question Answering

Open-domain question answering (ODQA) aims to answer factoid questions with a large collection of documents (Voorhees and Tice, 2000). With the new advances in large language models (LLMs), a typical approach to OPQA involves a two-stage framework: (1) first retrieve a small subset of passages where some of them contain the answer to the question, and then (2) use a LLM to answer the question using the retrieved passages as contexts (Chen et al., 2017; Karpukhin et al., 2020; Guu et al., 2020; Khandelwal et al., 2020; Izacard and Grave, 2021; Borgeaud et al., 2022; Zhong et al., 2022). Retrievers augment question answering by retrieving up to 100 passages and set new state-of-the-art for ODQA (Izacard and Grave, 2021); however, we believe that with such large amount of passages retrieved as context, it's frequent for them to contain conflicting information, and such conflicts will confuse the downstream language models in question answering. In this work, we validate this hypothesis and show that for a retriever like Google Search, 25% of the time it will return conflicting contexts in its top ten results when queried with a realistic, unambiguous question. We further demonstrate the limitations of current LLMs under this scenario of conflicting contexts through our experiments.

2.2 Knowledge Conflicts

Parametric v.s. Contextual One line of work studies knowledge conflicts in the setting of parametric v.s. contextual knowledge. Parametric knowledge refers to the knowledge a model learns during pre-training, and contextual knowledge refers to the contextual information a model sees at inference time. Longpre et al. (2021) proposes a entity-substitution framework that identifies QA instances with named entity answers and then substitutes mentions of the entity in the gold docu-

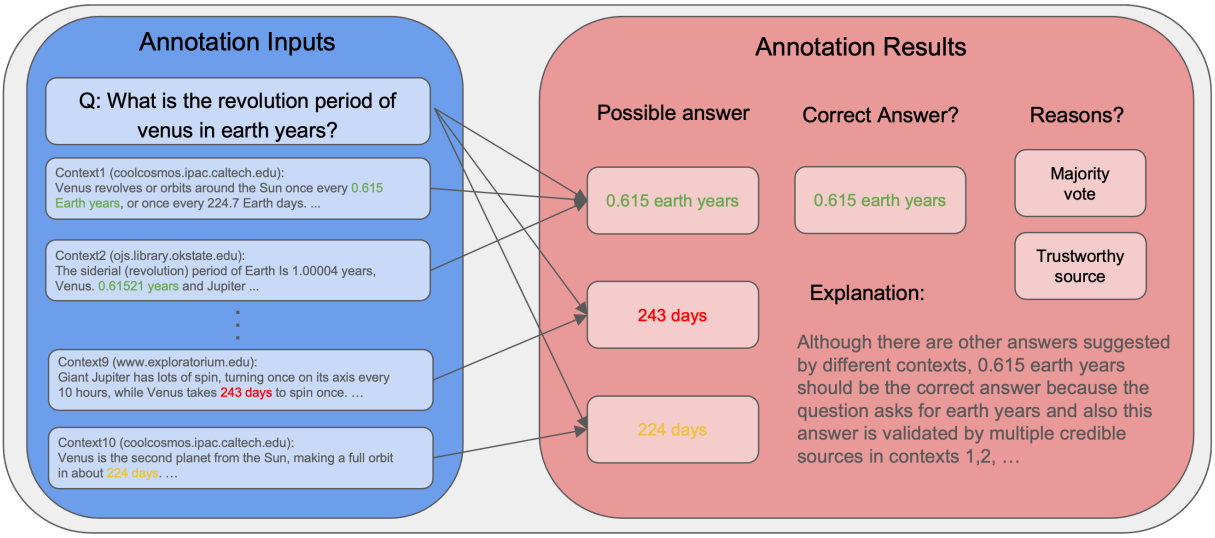


Figure 2: Data Collection Pipeline. The left side shows an example input given to the annotators, and the right side shows an example annotated result. During the annotation process, we ask the annotators to identify different possible answers given each context, and decide there is a conflict if there is more than one possible answer. In this example, the annotators believe there are three possible answers, and they think *0.615 earth years* is the correct answer because it's validated by most trustworthy sources.

ment with an alternate entity to create entity-based knowledge conflicts. Chen et al. (2022a) expands the study to consider multiple evidence passages and shows that when some passages are perturbed not to support an answer, language models largely ignore semantic perturbations and outputs potential answer entity in the retrieved passages. Xie et al. (2024) proposes another new framework to elicit the parametric memory of LLMs in order to construct the corresponding counter-memory and shows that with both supportive and contradictory evidence to their parametric memory, LLMs show a strong confirmation bias and tend to cling to their parametric memory. (Liu et al., 2024) proposes a machine-generated dataset of knowledge conflicts and studies different strategies to enable LLMs to resolve conflicting knowledge.

Contextual v.s. Contextual Another line of work focuses on the scenario when a language model is given conflicting contexts as in our setting. Some previous work create conflicting contexts with perturbations, including entity-substitution (Chen et al., 2022a; Hong et al., 2024) and machine-generation (Pan et al., 2023; Wan et al., 2024; Hong et al., 2024), and some other work define conflicts over rule-based templates (Kazemi et al., 2023). However, most of these previous work are built on synthetic data, whereas our dataset are real-world search results of unambiguous questions. Wan et al. (2024) also uses Google Search to extract conflicting contexts, but they focus specifically on controversial and contentious questions and analyze the linguistic features in the text that affect language models' predictions, whereas we show that realistic, unambiguous questions can also lead to conflicting contexts from Google and finetuning LLMs' on our human written explanations can teach them to reason through the conflicts.

3 Question Answering with Conflicting Contexts

In this section, we discuss our exploration of the problem: question answering with conflicting contexts. We first suggest our definition of what constitutes as conflicting context, then introduce how we collect a our dataset QACC for the analysis, and lastly share our findings and analysis of QACC.

3.1 Problem Definition

Given a question q, a list of retrieved contexts $C = \{c_1, c_2, ..., c_i\}$, and a question answering system ϕ, we can get a list of individual answers $A = \{a_1, a_2, ..., a_i\}$, where $a_i = \phi(q, c_i)$. We state that the question q has conflicting evidence if and only if $\exists(a_i, a_j \in A)(a_i \neq a_j)$. In other words, at each step a question answering system (a human or a language model) is given the question and only one of the context in order to answer the question. Iterate this through all of the contexts, and we state that there are conflicting contexts if

Split	# of QAs	% QAs with Conflicts
Train	394	29%
Dev	303	19%
Test	813	25%

Table 1: The statistics of QACC.

and only if there are different answers generated when given different contexts. Note that our definition of *conflict* here differs from the definition of *contradiction* in traditional Natural Language Inference (NLI) tasks. Here we define *conflict* in a way that is less restricting than *contradictory* texts in NLI to further exploit its applicability in our target domain, open domain QA with RAG. For instance, in a RAG scenario, retrievers can often retrieve contexts that contain seemingly correct answers (e.g. *neutral* texts in the case of NLI), and such different/conflicting answers may also confuse the downstream LLMs. We believe that our definition of *conflict* and the resulted dataset can therefore better support us towards our goal.

3.2 QACC Dataset

Ambiguous questions can frequently lead to multiple different answers (Min et al., 2020). However, we believe that even when questions are *unambiguous*, it is still common to see conflicting evidence on the web. To this end, we consider AmbigQA (Min et al., 2020), a dataset with questions labeled as either *ambiguous* or *unambiguous*, and take only questions that are labeled as *unambiguous* as the the questions in our dataset. We then use Google Search API to retrieve top-10 search results as the contexts for each question, and use Amazon Mechanical Turk to collect annotations for each question and its associated contexts. The statistics of our dataset QACC is shown in Table 1.

3.3 Human Annotation

We employ a rigorous human annotation pipeline with a qualification exam before the main annotation task, a strategy commonly used to ensure the collection of high-quality datasets (Han et al., 2021; Dasigi et al., 2021). Only annotators that have passed our qualification exams can participate in the main annotation task. We use Amazon Mechanical Turk to collect the annotations and CROWDAQ to design the annotation interface(Ning et al., 2020). Each question is annotated by one annotator and paid with $0.35 in the main annotation task. Examples of our qualification and annotation interfaces are shown in Appendix D.

Qualification Exam Since the annotation task requires critical thinking and an attention to detail, we design interactive tutorials and request the annotators to review them before the qualification exams. We first show them instructions and our definition of conflicting contexts for a question, and then ask them to complete a set of tutorial questions where we display the expected answers and reasons once they answer them. After they understand the goals and formats of the annotations, we request them to complete a set of 12 random, multiple-choice qualification questions. Only workers with more than 90% accuracy on the exam can pass and get the qualification to participate in our main annotation task. We allow only the workers that have a master's qualification[3] to take the exam, and 12 among 41 of them (29%) have passed our exam and participate in our main annotation.

Main Annotation Following our definition of the problem, we ask the annotators to identify the conflict in the contexts by finding different possible answers. In each Human Intelligence Task (HIT), we show the annotator an open domain question, a list of contexts retrieved by Google Search, as well as the website domains these contexts are from. We then ask the annotators the following questions: 1. are there more than one possible answer when looking at the question and each context individually (conflict identification)? 2. which of the contexts support which of the different answers (answer attribution)? 3. which answer do you think is the correct answer (question answering)? 4. why do you think the answer you choose is correct (QA with explanation)?. The fourth question here includes both a multiple-choice question that asks them to select a reason from a pre-defined set and a free-form question that asks them to explains their reasoning in a single sentence. These procedures result in a rich annotation of QACC that can also support other QA-related tasks not covered in the scope of this work, like answer attribution. An example of the dataset and the data collection process is shown in Figure 10.

3.4 Expert Verification

To verify the quality of the annotations, we take a random sample of 20 examples from QACC and

[3]Amazon Mechanical Turk award workers master's qualifications only if they have demonstrated superior performance over a period of time across thousands of annotations.

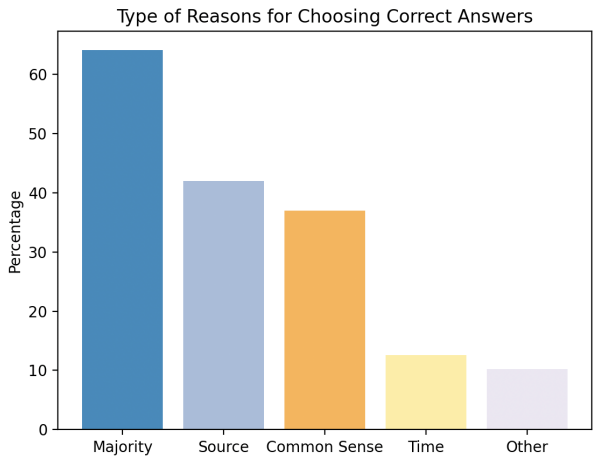


Figure 3: Reasons of annotators selecting one correct answer over the others when there are conflicts. "Majority" means the answer is supported by the most contexts. "Source" means the annotator trust the contexts more because they come from trustworthy sources. "Common Sense" means the answer matches their own memory and common sense. "Time" means they think one answer is correct since it's the most up-to-date.

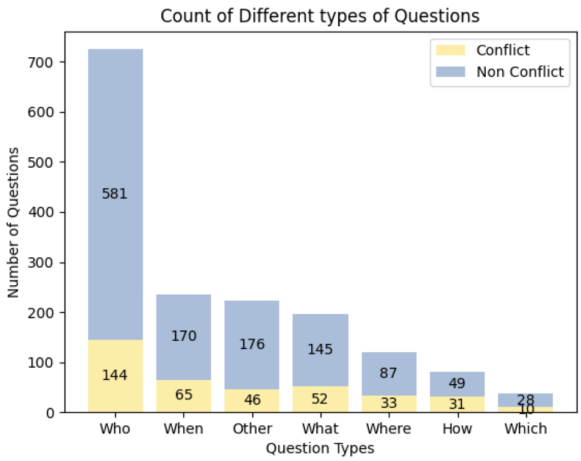


Figure 4: Different types of questions in our dataset that have conflicts.

annotate whether there is a conflict in the contexts ourselves. We match our annotation with the Mechanical Turk Workers' annotations and observe a Cohen's kappa agreement of 0.9, showing a high quality of the annotations in our dataset.

3.5 Analysis

Table 1 shows the statistics of our dataset. We can see that about 25% of all the unambiguous, open domain questions in our dataset have conflicting contexts when retrieved using Google Search.

To better understand humans' reasoning process when presented with conflicting evidence, we ask the annotators to choose from a pre-defined list of reasons that can best categorize why they think one of the answers is correct. We allow them to choose more than one option since different factors can simultaneously affect one's decision in choosing the correct answer. Figure 3 shows their reasons when the question is labeled by them as having conflicting contexts. We find that humans favor answers that are the most popular in the contexts the most, and also refer to the sources of the context (trustworthy or not) and their own intuitions and common sense about the question when deciding on the answer. On the other hand, less annotators select correct answers based on the publish time of the information.

We also conduct data analysis to study the different types of questions in our dataset that lead to conflicting contexts. In figure 4, we see that most of the questions in our dataset are *Who* questions. This type of question has about 20% of times that lead to conflicting contexts. *How* questions, on the other hand, can lead to conflicting contexts almost 40% of the times. This aligns with our hypothesis since questions that start with *How* are typically open-ended questions with a more complex answer and can involve different perspectives.

Another characteristic about QACC that we observe is that the most prominent type of conflicts between different answers is the mismatch of entities, i.e., different names, time, and places. This is largely due to the fact that we source our questions from AmbigQA and Natural Questions, where the expected answers are short phrases that are mostly entities.

4 Experiments

In this section, we benchmark the problem of question answering with conflicting contexts on our dataset with three popular LLMs. We demonstrate that teaching language models to explain its answer can guide their inference process and improve their performance in QACC, and such improvement can generalize to another perturbed NQ-Open dataset.

4.1 Datasets

We run our experiments on two datasets. The first is the QACC dataset we collect, and the second is a perturbed NQ-Open dataset. For our QACC dataset, we use the validation set to find the best instruction and prompt formats for the LLMs, and report their

Model	Prompt	EM-C	EM-NC	EM-T	F1-C	F1-NC	F1-T
GPT-4o-fewshot	Context	42.51	54.13	51.17	59.29	**70.06**	**67.32**
	+ Majority & Ans	40.58	53.96	50.55	57.35	69.03	66.05
	+ Discern & Ans	39.13	52.15	48.83	56.33	67.81	64.88
	+ Exp & Ans	38.65	54.29	50.31	57.43	69.19	66.2
Claude-3-fewshot	Context	35.75	53.14	48.71	53.76	67.9	64.3
	+ Majority Vote	37.68	52.64	48.83	54.75	67.85	64.51
	+ Dis & Ans	31.4	43.89	40.71	48.52	58.48	55.95
	+ Exp & Ans	40.1	54.29	50.68	57.41	69.28	66.26
Phi-3-fewshot	Context	42.51	51.98	49.57	56.65	67.30	64.59
	+ Majority Vote	**44.93**	55.12	52.52	58.29	69.18	66.41
	+ Dis & Ans	43.96	54.79	52.03	58.41	68.51	65.94
	+ Exp & Ans	43.48	**55.78**	**52.64**	**59.57**	69.32	66.84
Phi-3-finetuned	Context	44.44	51.32	49.57	56.24	64.6	62.47
	+ Majority Vote	44.44	55.94	53.01	57.26	67.94	65.22
	+ Dis & Ans	38.16	49.01	46.25	49.99	61.86	58.84
	+ Exp & Ans	**47.34**	**57.26**	**54.74**	**59.61**	**69.79**	**67.19**

Table 2: Fewshot and finetuned results of models and methods tested on our QACC dataset. EM-C means the Exact Match (EM) score of the set of QA pairs with conflicting contexts, EM-NC means the EM score of QAs with non-conflicting contexts, and EM-T means the total EM score of all the QAs in the test set. Same notation applies to the F1 score. "Context" means LLMs are given both the question and contexts retrieved from google. "+ Majority Vote" means LLMs are given question, contexts and the instruction to take majority vote. "+ Dis & Ans" indicates LLMs are given question, contexts, the instruction to discern and answer, and either an in-context example or finetuning data indicating which contexts are perturbed. "+ Exp & Ans" represents results of LLMs with question, contexts, the instruction to explain and answer, and in-context example of explanation or finetuning data of explanation. The **bolded** numbers represent the best results across all fewshot models or finetuned models, and the underlined numbers represent the best result in a single model.

results on the test set. For the perturbed NQ-Open dataset, we use an entity-substitution method to replace the answer in the contexts to other named entities of the same type in order to create conflicts among the contexts, following (Longpre et al., 2021). We construct this perturbed dataset over the test split of NQ-Open with 3,610 questions. We retrieve top ten results from Google as the contexts for these questions and apply the entity-substitution algorithm with different perturbation ratio. The higher the perturbation ratio means the more contexts in a question are perturbed.

4.2 Methods

Retrieval Augmented QA Language models can leverage contexts to answer open domain questions (Izacard and Grave, 2021; Zhong et al., 2022). We prompt LLMs with question and contexts retrieved from Google and instruct them that *"Given the following contexts, provide a direct, concise answer to the question"*.

Majority Vote As shown in Figure 3, humans are inclined to choose the majority answer when there are conflicting evidence. Therefore, we prompt LLMs question and contexts and instruct them to *"use majority vote to decide which context to trust if there are conflicting contexts"*.

Discern and Answer Hong et al. (2024) proposes to explicitly instruct the model to first discern the counterfactual, perturbed passages and then ignore them to answer the question. We follow the same strategy and instruct the models to *"Find the perturbed passages if there are any, and ignore them when eliciting the correct answer"* with question, contexts, and an example message indicating which of the contexts are "perturbed", using the annotations in QACC that attribute correct/wrong answers to their supporting contexts.

Explain and Answer Prompting with explanations introduces richer information that can guide the inference process. Recent work have shown that letting the language model "explain itself" through in-context learning gains more insights into predictions and improves their performances in a variety of reasoning tasks, including question answering

Model	Prompt	Perturbation Ratio					
		0 %		25 %		50 %	
		EM	F1	EM	F1	EM	F1
Phi-3	Zeroshot Context	15.60	33.08	14.49	31.68	13.21	29.81
	Fewshot Context	30.25	44.21	28.59	42.40	25.93	39.84
Phi-3-finetuned	Context	34.24	47.22	32.27	45.34	29.67	42.80
	Majority Vote	34.29	46.67	32.71	45.07	29.72	42.31
	Dis & Ans	32.44	44.26	30.61	42.60	28.22	40.16
	Exp & Ans	**37.31**	**50.29**	**35.76**	**48.57**	**33.18**	**46.20**

Table 3: Zeroshot, fewshot and finetuned results on perturbed NQ-Open test set. The higher the perturbation ratio, the more contexts in a question are perturbed with entity-substitution.

(Lampinen et al., 2022b; Ye and Durrett, 2022; Nye et al., 2021; Wei et al., 2023; Lampinen et al., 2022a). We believe answering question with conflicting contexts requires similar reasoning abilities and therefore can benefit from eliciting explanations during inference. We instruct the models to *"Explain the reasons and then provide a direct, concise answer"* with question, contexts, and a natural language explanation as in-context example in fewshot and training input in finetuning. Table 4 shows our experiments of comparing Explain then Answer to Answer then Explain. Similar to previous work, we observe only slight impact of the orders of explanation in their performances.

Model	EM	F1
Fewshot Exp & Ans	52.64	66.84
Finetuned Exp & Ans	**54.74**	67.19
Fewshot Ans & Exp	51.41	65.78
Finetuend Ans & Exp	53.75	**67.24**

Table 4: Experiments on the order of explanation and answer on Phi-3-Medium evaluated on QACC. We observe only slight differences in terms of the performance.

4.3 Experiment Setup

We conduct experiments on three popular instruction-tuned large language models from different families: *GPT-4o-mini*, *Claude3-Sonnet*, and *Phi-3-Medium-Instruct (14B)*, with zero-shot inference, few-shot inference, and finetuning. We find that LLMs greatly benefit from in-context examples (few-shot) compared to zeroshot (see Appendix B) when answering open domain questions, so we only present few-shot inference and finetuning results in Table 2.

For few-shot inference experiments, we include **one** in-context example of expected input-output pair when prompting the three language models. For finetuning experiments, we finetune Phi-3-Medium-Instruct using LoRA (Hu et al., 2021) with language modeling loss (SFT). We first find the best hyperparameters of finetuning using the validation set of QACC and then train on both the training and validation set and report results in the test set. We also use the validation set of QACC to find the best prompt and instruction format for each methods and use them for both fewshot inference and finetuning. More details of experiment settings are discussed in Appendix A

We follow the conventions and use Exact Match and F1 scores as the metrics for all our evaluations. Exact Match returns positive if the generated answer is identical to the reference and negative if otherwise, whereas F1 score is more forgiving and measures the word overlap between the generated and reference answers. We note that LLMs are prone to long generations, so we specifically instruct all of the models to *answer with as few words as possible* in the prompts (see examples of the prompts in Appendix C).

4.4 Experiment Results

QACC Table 2 exhibits our experiment results in QACC. We can see that all LLMs that we evaluate inevitably experience worse performance when there are conflicting contexts, comparing their results on EM-C and EM-NC, as well as their results on F1-C and F1-NC. We also find that in different LLMs, their best prompting methods in the few-shot setting are also different. GPT-4o has the best performance when prompted with just the contexts when seeing conflicting contexts, Claude-3 gives the best results when instructed to first explain and then answer the question, and Phi-3 presents comparable performances when instructed to take ma-

jority vote and explain then answer.

We also demonstrate that by instructing the model to explain its answer and finetuning with our human-written explanations, Phi-3 can improve its performance on question answering with conflicting contexts. We observe an improvement of 2.9% on EM and 3.37% on F1 comparing the models finetuned with just the contexts (Context) and with contexts and the explanations (+Exp & Ans). Interestingly, we find that by finetuning Phi-3 with contexts and the instruction to take Majority Vote, the model cannot further improve its performance, and finetuning with Discern and Answer instruction and examples hurts the model and diminishes its performance. We hypothesize the reason is that by finetuning with the instruction to take Majority Vote, we are not introducing any new learning signals to the models besides the format, which it already learns from in-context examples, and some QA examples, which Phi-3 may have already seen during its pre-training. On the other hand, finetuning with Discern and Answer data hurts the performance since, although we can attribute the answers to their supporting contexts to create finetuning data for it, our conflicting contexts are naturally existed conflicting information on the web, rather than synthetic perturbed data with only a few entities replaced. This discrepancy re-emphasizes the usefulness of our dataset with naturally conflicting contexts.

Perturbed NQ-Open The improvements we observe from finetuning Phi-3 on human-written explanations can generalize to perturbed data and general open-domain question answering as well. Table 3 exhibits the performance of zeroshot, fewshot Phi-3 models on perturbed NQ-Open, as well as Phi-3 finetuned on QACC and evaluated on perturbed NQ-Open. As illustrated in Table 3, Phi-3 finetuned with explanation data consistently outperforms other finetuned models under different ratio of perturbation. We believe that the extra finetuning signals of natural language explanations improve Phi-3's reasoning abilities in general, therefore demonstrating its consistent improvements across all perturbation ratio, including regular open domain QA (0%) and when there are perturbed contexts (25% and 50%).

4.5 Analysis

Figure 5 shows the performance of fewshot GPT-4o on the set of questions in QACC that has conflicting contexts. We can see that LLMs like GPT-4o can answer open domain questions reasonably well even with conflicting contexts when the questions are asking about a person (*Who* types of question). However, GPT-4o fails significantly more when the question is asking about a time (*When*), about a place (*Where*), or when the question is open-ended (*How*). We hypothesize that this may relate to the pretraining corpus of LLMs and the frequency that different entities appear in the pretraining corpus: popular people names that appear frequently in LLMs' pretraining corpus allow them to utilize their parametric knowledge to distinguish the answers among the conflicting contexts, whereas the different timestamps and places exist more sparsely in the corpus (as well as on the web), making it more difficult for the LLMs to discern.

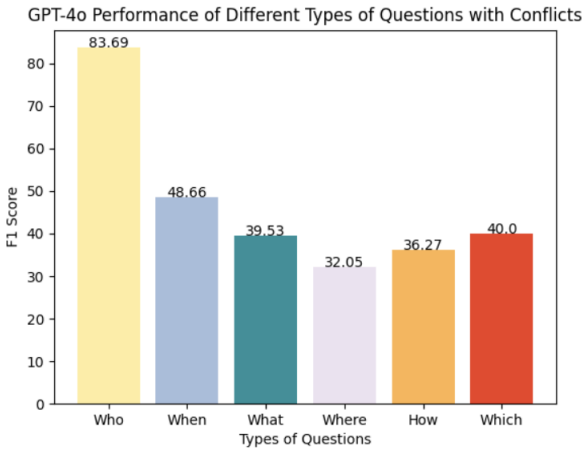


Figure 5: Fewshot GPT-4o performance on the test set of QACC that has conflicting contexts. The x-axis indicates the different types of questions and the y-axis denotes the F1 score for each type.

5 Conclusion

In this work, we construct a dataset named QACC to study open domain question answering with conflicting contexts. We find that unambiguous, open domain questions are exposed to conflicting evidence on the web: 25% of the questions will lead to conflicting contexts when retrieved using Google, and popular LLMs are very brittle to such conflicts. We show that by finetuning on natural language explanations, we can improve the reasoning abilities of Phi-3 and improve its performances when there are conflicting contexts as well as open domain question answering in general. We will release our dataset and code to promote further research along this line.

Limitations Eliciting natural language explanations from LLMs have several limitations. Previous work have shown that explanations generated by LLMs can be unreliable and can lead to wrong interpretations of the models. However, in this work, we focus on the improvement of reasoning abilities of LLMs when finetuning with explanations data, rather than making interpretations of their explanations.

References

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. *Preprint*, arXiv:2112.04426.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. *Preprint*, arXiv:1704.00051.

Hung-Ting Chen, Michael Zhang, and Eunsol Choi. 2022a. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2292–2307, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing things from a different angle: Discovering diverse perspectives about claims. *arXiv preprint arXiv:1906.03538*.

Sihao Chen, Siyi Liu, Xander Uyttendaele, Yi Zhang, William Bruno, and Dan Roth. 2022b. Design challenges for a multi-perspective search engine. *Preprint*, arXiv:2112.08357.

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. *Preprint*, arXiv:2105.03011.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *Preprint*, arXiv:2002.08909.

Katri Hämeen-Anttila, Hedvig Nordeng, Esa Kokki, Johanna Jyrkkä, Angela Lupattelli, Kirsti Vainio, and Hannes Enlund. 2014. Multiple information sources and consequences of conflicting information about medicine use during pregnancy: A multinational internet-based survey. *J Med Internet Res*, 16(2):e60.

Rujun Han, I-Hung Hsu, Jiao Sun, Julia Baylon, Qiang Ning, Dan Roth, and Nanyun Peng. 2021. ESTER: A machine reading comprehension dataset for reasoning about event semantic relations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7543–7559, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Giwon Hong, Jeonghwan Kim, Junmo Kang, Sung-Hyon Myaeng, and Joyce Jiyoung Whang. 2024. Why so gullible? enhancing the robustness of retrieval-augmented models against counterfactual noise. *Preprint*, arXiv:2305.01579.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. *Preprint*, arXiv:2007.01282.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense passage retrieval for open-domain question answering. *Preprint*, arXiv:2004.04906.

Jungo Kasai, Keisuke Sakaguchi, yoichi takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, and Kentaro Inui. 2023. Realtime qa: What's the answer right now? In *Advances in Neural Information Processing Systems*, volume 36, pages 49025–49043. Curran Associates, Inc.

Mehran Kazemi, Quan Yuan, Deepti Bhatia, Najoung Kim, Xin Xu, Vaiva Imbrasaite, and Deepak Ramachandran. 2023. Boardgameqa: A dataset for natural language reasoning with contradictory information. *Preprint*, arXiv:2306.07934.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through memorization: Nearest neighbor language models. *Preprint*, arXiv:1911.00172.

Andrew K. Lampinen, Ishita Dasgupta, Stephanie C. Y. Chan, Kory Matthewson, Michael Henry Tessler, Antonia Creswell, James L. McClelland, Jane X. Wang, and Felix Hill. 2022a. Can language models learn from explanations in context? *Preprint*, arXiv:2204.02329.

Andrew K. Lampinen, Nicholas A. Roy, Ishita Dasgupta, Stephanie C. Y. Chan, Allison C. Tam, James L. McClelland, Chen Yan, Adam Santoro, Neil C. Rabinowitz, Jane X. Wang, and Felix Hill. 2022b. Tell me why! explanations support learning relational and causal structure. *Preprint*, arXiv:2112.03753.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open

domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *Preprint*, arXiv:2307.03172.

Siyi Liu, Sihao Chen, Xander Uyttendaele, and Dan Roth. 2021. Multioped: A corpus of multi-perspective news editorials. *arXiv preprint arXiv:2106.02725*.

Yantao Liu, Zijun Yao, Xin Lv, Yuchen Fan, Shulin Cao, Jifan Yu, Lei Hou, and Juanzi Li. 2024. Untangle the knot: Interweaving conflicting knowledge and reasoning skills in large language models. *Preprint*, arXiv:2404.03577.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. Ambigqa: Answering ambiguous open-domain questions. *Preprint*, arXiv:2004.10645.

Qiang Ning, Hao Wu, Pradeep Dasigi, Dheeru Dua, Matt Gardner, Robert L. Logan IV, Ana Marasović, and Zhen Nie. 2020. Easy, reproducible and quality-controlled data collection with CROWDAQ. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 127–134, Online. Association for Computational Linguistics.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. 2021. Show your work: Scratchpads for intermediate computation with language models. *Preprint*, arXiv:2112.00114.

Liangming Pan, Wenhu Chen, Min-Yen Kan, and William Yang Wang. 2023. Attacking open-domain question answering by injecting misinformation. *Preprint*, arXiv:2110.07803.

Ellen M. Voorhees and Dawn M. Tice. 2000. The TREC-8 question answering track. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Alexander Wan, Eric Wallace, and Dan Klein. 2024. What evidence do language models find convincing? *Preprint*, arXiv:2402.11782.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. *Preprint*, arXiv:2305.13300.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. *Preprint*, arXiv:2205.03401.

Michael J. Q. Zhang and Eunsol Choi. 2021. Situatedqa: Incorporating extra-linguistic contexts into qa. *Preprint*, arXiv:2109.06157.

Zexuan Zhong, Tao Lei, and Danqi Chen. 2022. Training language models with memory augmentation. *Preprint*, arXiv:2205.12674.

A Experiment Details

We perform our finetuning experiments using Amazon Elastic Compute Cloud (Amazon EC2). We use one p4d.24xlarge instance for the training. It has 8 NVIDIA A100 GPUs with 40.0 GB GPU memory each. We use Parameter-Efficient Fine-Tuning (PEFT) and Low-Rank Adaptation (LoRA) to train our models. We use learning rata 2e-4, lora_r = 16, lora_alpha = 32, and lora_dropout = 0.05.

We use snippets returned by Google Search as the contexts for each of the articles retrieved. Studies have shown that LLMs struggle to process and understand very long contexts (Liu et al., 2023), so we believe that Google snippets are good summarization and extraction of their original articles for our model to process different contexts. To guarantee that the answer always exists in the contexts, we remove examples that our annotators believe they didn't find a correct answer in the contexts in our experiments. Examples of our annotation interface are shown in Appendix D.

B Zeroshot v.s. Fewshot Inference

As shown in Table 5, teaching large language models (LLMs) through few-shot examples can effectively demonstrate to them the format and expected output of the task and improve performance significantly.

C Example Prompt

We present an example of the prompt and in-context example we use for running our experi-

Model	Instruction	EM	F1
Phi3-Mini	0-SHOT-CONTEXT	32.67	54.95
Phi3-Mini	1-SHOT-CONTEXT	47.6	63.51
Phi3-Medium	0-SHOT-CONTEXT	45.14	63.45
Phi3-Medium	1-SHOT-CONTEXT	**49.57**	**64.59**

Table 5: Zeroshot v.s. fewshot for Phi-3 on QACC.

ments. We preserve the same format and instruction in our finetuning experiments for consistency.

Instruction

```
Given the following contexts, explain your reasoning first and then provide a direct, concise answer to the question.
Make the answer as short as possible and use \ n \ n as the delimiter to distinguish your explanation and answer.
```

In-context example

```
Context1 from www.sho.com: Kerris Lilla Dorsey is best known for her roles as Brad Pitt's daughter in the
Oscar® nominated film MONEYBALL and as Steve Carell and Jennifer Garner's ...
Context2 from www.imdb.com: She is known for her roles as Paige Whedon in the television series Brothers &
Sisters, Casey Beane, Billy Beane's (Brad Pitt) daughter, in the 2011 film ...
Context3 from people.com: Jul 2, 2015 ... Kerris Dorsey has worked with several of Hollywood's most
famous dads. She currently plays Liev Schreiber's daughter on Ray Donovan and got ...
Context4 from www.imdb.com: She is known for her roles as Paige Whedon in the television series Brothers & Sisters,
Casey Beane, Billy Beane's (Brad Pitt) daughter, in the 2011 film ...
Context5 from www.sportskeeda.com: Feb 11, 2023 ... All about Casey Beane, who was pictured in Moneyball as a young
girl ... The 2011 film Moneyball was one of the most popular baseball movies ever ...
Context6 from en.wikipedia.org: Moneyball is a 2011 American biographical sports drama film directed by Bennett
Miller with ... In the film, Beane (Brad Pitt) and assistant general manager Peter Brand ...
Context7 from www.latimes.com: Sep 10, 2011 ... And “Moneyball” is all about Brad Pitt. ... And while you get
a couple of glimpses of Pitt's daughter, played by 13-year-old Kerris Dorsey, ...
Context8 from www.pinterest.com: Jan 9, 2020 ... ... (played by Brad Pitt) daughter, in the 2011 film Moneyball,
... Dorsey plays Bridget Donovan, the daughter of the title character, ...

Who plays brad pitt's daughter in moneyball?

Explanation: I chose it because it is the only correct answer and it is explicitly mentioned in context 4 and 5.

Answer: Kerris Lilla Dorsey
```

D Human Annotation Interface

Question

who played spock's dad on star trek

Context 1

Lenard is best known for his appearances in the Star Trek franchise, particularly in the role of Sarek, the father of Spock (Leonard Nimoy).

Context 2

Nov 28, 1996 ... Mark Lenard, an actor who performed Shakespeare, Chekhov and Ibsen but was best known for playing Mr. Spock's Vulcan father in the “Star ...

Context 3

Aug 18, 2020 ... Star Trek fans will best know Cross as the actor who played Spock's father Sarek in J.J. Abrams' 2009 Star Trek reboot.

Context 4

Aug 19, 2020 ... Cross played Spock's father Sarek in the 2009 reboot of “Star Trek” and portrayed Prince Charles in the television film “William & Kate: The ...

Context 5

Nov 11, 2020 ... The following season, Lenard played a visually similar if characteristically different alien altogether. He was Spock's dad, Sarek, in "Journey ...

Context 6

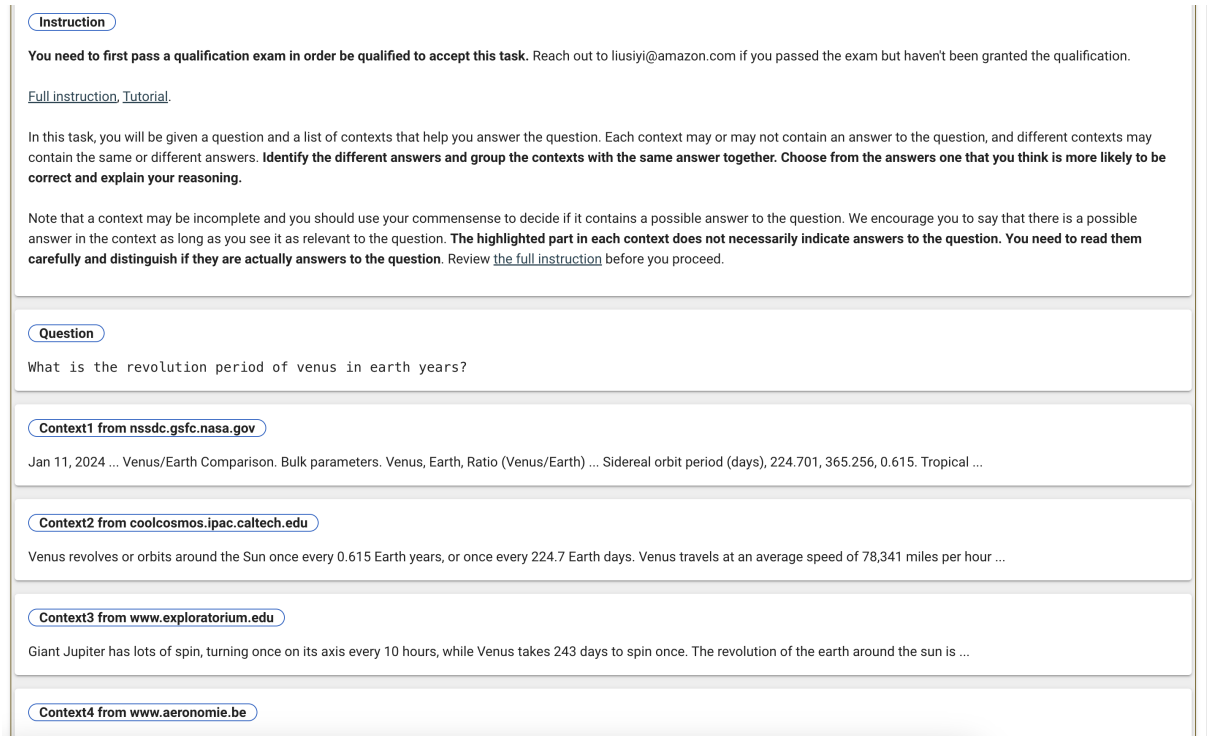
Spock's father was played by actor Mark Lenard in the original Star Trek series and in several of the films. Continue reading.

Figure 6: Qualification example 1.

Remember that we define conflicts as having different, possible answers in the different contexts. Are there conflicts in the contexts above?

- Yes, there are conflicts in the contexts.
- No, there are no conflicts in the contexts.

Figure 7: Qualification example 1.



Instruction

You need to first pass a qualification exam in order be qualified to accept this task. Reach out to liusiyi@amazon.com if you passed the exam but haven't been granted the qualification.

Full instruction, Tutorial.

In this task, you will be given a question and a list of contexts that help you answer the question. Each context may or may not contain an answer to the question, and different contexts may contain the same or different answers. **Identify the different answers and group the contexts with the same answer together. Choose from the answers one that you think is more likely to be correct and explain your reasoning.**

Note that a context may be incomplete and you should use your commensense to decide if it contains a possible answer to the question. We encourage you to say that there is a possible answer in the context as long as you see it as relevant to the question. **The highlighted part in each context does not necessarily indicate answers to the question. You need to read them carefully and distinguish if they are actually answers to the question**. Review the full instruction before you proceed.

Question

What is the revolution period of venus in earth years?

Context1 from nssdc.gsfc.nasa.gov

Jan 11, 2024 ... Venus/Earth Comparison. Bulk parameters. Venus, Earth, Ratio (Venus/Earth) ... Sidereal orbit period (days), 224.701, 365.256, 0.615. Tropical ...

Context2 from coolcosmos.ipac.caltech.edu

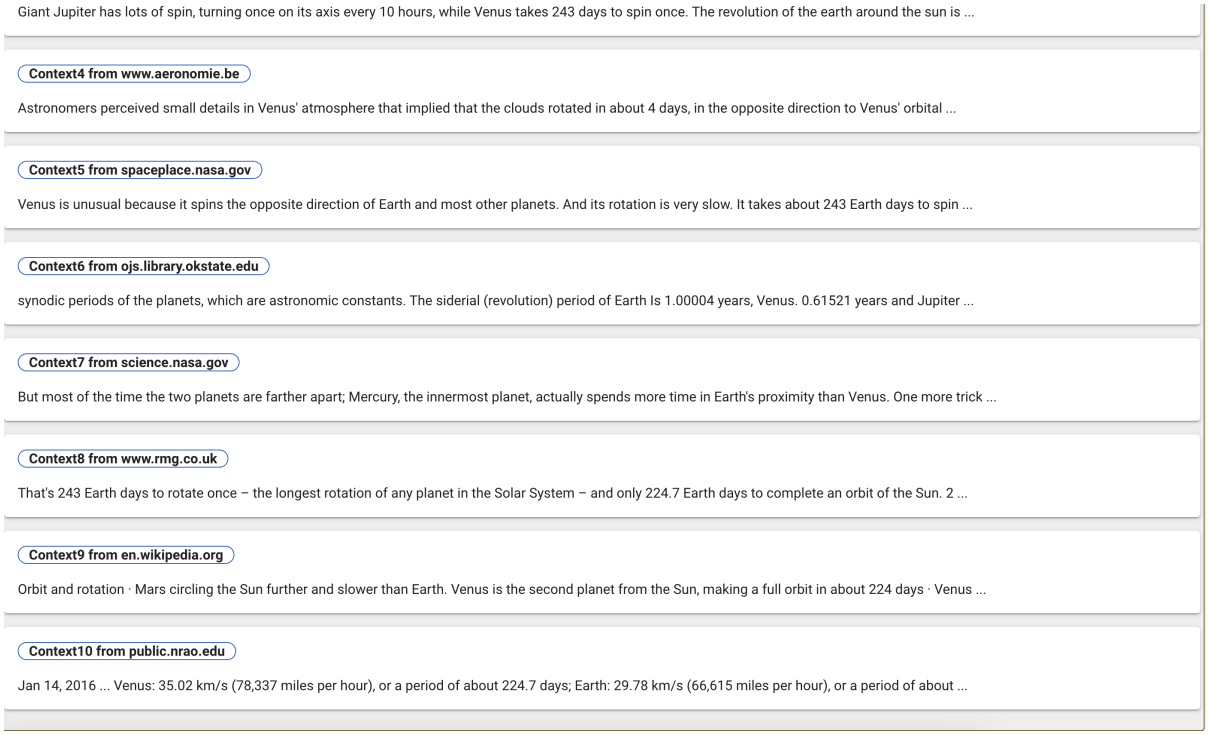
Venus revolves or orbits around the Sun once every 0.615 Earth years, or once every 224.7 Earth days. Venus travels at an average speed of 78,341 miles per hour ...

Context3 from www.exploratorium.edu

Giant Jupiter has lots of spin, turning once on its axis every 10 hours, while Venus takes 243 days to spin once. The revolution of the earth around the sun is ...

Context4 from www.aeronomie.be

Figure 8: Annotation example 1.



Giant Jupiter has lots of spin, turning once on its axis every 10 hours, while Venus takes 243 days to spin once. The revolution of the earth around the sun is ...

Context4 from www.aeronomie.be

Astronomers perceived small details in Venus' atmosphere that implied that the clouds rotated in about 4 days, in the opposite direction to Venus' orbital ...

Context5 from spaceplace.nasa.gov

Venus is unusual because it spins the opposite direction of Earth and most other planets. And its rotation is very slow. It takes about 243 Earth days to spin ...

Context6 from ojs.library.okstate.edu

synodic periods of the planets, which are astronomic constants. The siderial (revolution) period of Earth Is 1.00004 years, Venus. 0.61521 years and Jupiter ...

Context7 from science.nasa.gov

But most of the time the two planets are farther apart; Mercury, the innermost planet, actually spends more time in Earth's proximity than Venus. One more trick ...

Context8 from www.rmg.co.uk

That's 243 Earth days to rotate once – the longest rotation of any planet in the Solar System – and only 224.7 Earth days to complete an orbit of the Sun. 2 ...

Context9 from en.wikipedia.org

Orbit and rotation · Mars circling the Sun further and slower than Earth. Venus is the second planet from the Sun, making a full orbit in about 224 days · Venus ...

Context10 from public.nrao.edu

Jan 14, 2016 ... Venus: 35.02 km/s (78,337 miles per hour), or a period of about 224.7 days; Earth: 29.78 km/s (66,615 miles per hour), or a period of about ...

Figure 9: Annotation example 2.

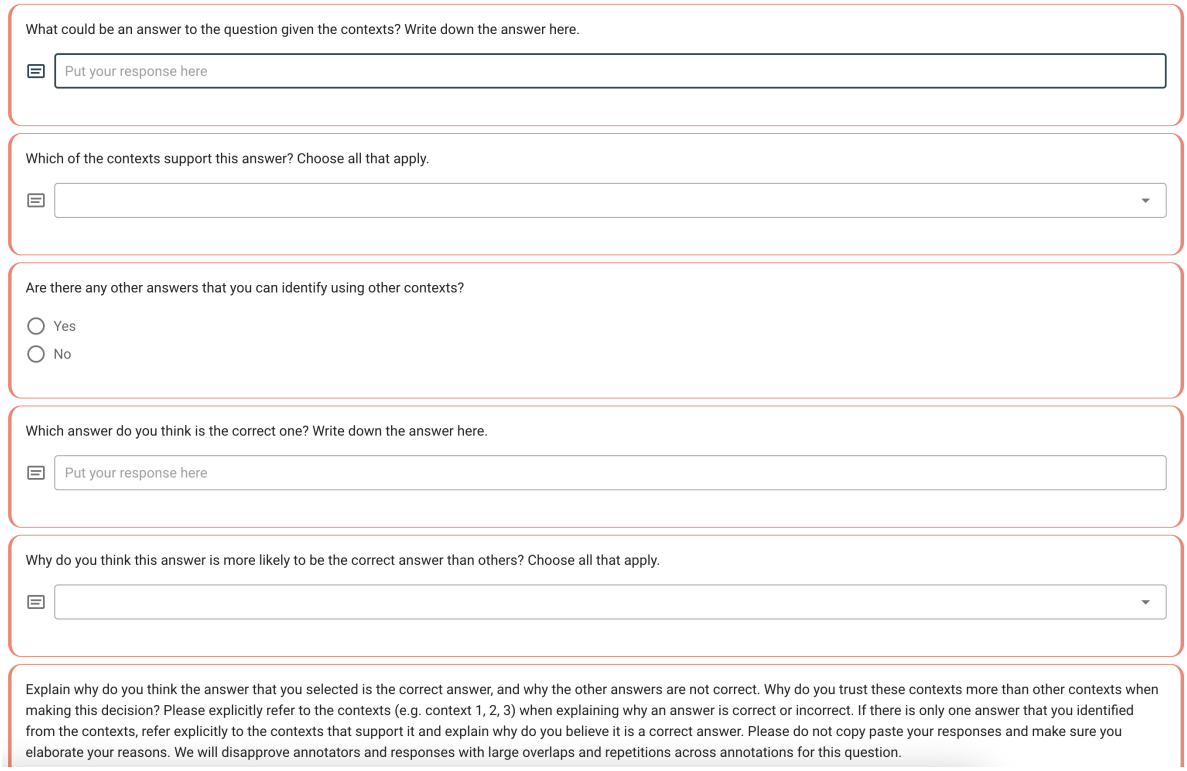


Figure 10: Annotation example 3.