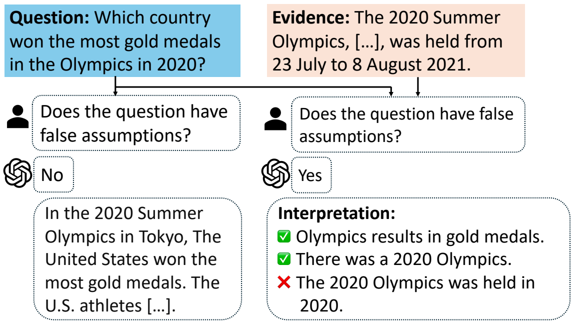
Identifying and Answering Questions with False Assumptions: An Interpretable Approach

Zijie Wang and **Eduardo Blanco**
Department of Computer Science
University of Arizona
{zijiewang,eduardoblanco}@arizona.edu

Abstract

People often ask questions with false assumptions, a type of question that does not have regular answers. Answering such questions requires first identifying the false assumptions. Large Language Models (LLMs) often generate misleading answers to these questions because of hallucinations. In this paper, we focus on identifying and answering questions with false assumptions in several domains. We first investigate whether the problem reduces to fact verification. Then, we present an approach leveraging external evidence to mitigate hallucinations. Experiments with five LLMs demonstrate that (1) incorporating retrieved evidence is beneficial and (2) generating and validating atomic assumptions yields more improvements and provides an interpretable answer by pinpointing the false assumptions.



Question: Which country won the most gold medals in the Olympics in 2020?

Evidence: The 2020 Summer Olympics, [...], was held from 23 July to 8 August 2021.

Does the question have false assumptions?

No

In the 2020 Summer Olympics in Tokyo, The United States won the most gold medals. The U.S. athletes [...].

Does the question have false assumptions?

Yes

Interpretation:
✅ Olympics results in gold medals.
✅ There was a 2020 Olympics.
❌ The 2020 Olympics was held in 2020.

Figure 1: Question with a false assumption (i.e., the 2020 Olympics was held in 2020). ChatGPT provides a misleading answer without identifying the false assumption. Our approach couples external evidence with a process to generate and validate atomic assumptions. Doing so allows us to answer questions with false assumptions by pointing out the false assumptions (e.g., There was a 2020 Olympics but it was not held in 2020).

1 Introduction

Large Language Models (LLMs) (Brown et al., 2020; Team et al., 2023) have demonstrated remarkable abilities in extractive (Rajpurkar et al., 2016; Kwiatkowski et al., 2019) and generative question answering (Reddy et al., 2019; Fan et al., 2019) among others. Despite their capabilities, LLMs suffer from hallucinations (Zhang et al., 2023). This leads to unfaithful answers due to overconfidence—when LLMs lack knowledge, they often make up answers despite abstaining (e.g., “I do not know”) being more desirable (Feng et al., 2024). Overconfidence also leads to hallucinated answers to unanswerable questions (Slobodkin et al., 2023). For example, “Which countries border Kansas?” should be addressed by pointing out that Kansas does not have an international border.

Unanswerable questions, as studied in previous works (Kim et al., 2023; Yu et al., 2023b; Hu et al., 2023), are information-seeking questions containing false assumptions. The false assumptions make them lack regular answers. Instead, answers should point out the false assumptions. Consider the question in Figure 1. It wrongly assumes that the 2020 Olympics was held in 2020. The event, however, was held in 2021. ChatGPT fails to identify the false assumption and generates a misleading answer. In this paper, we investigate the problem of identifying and answering questions with false assumptions. We define false assumptions in questions as subjective opinions, beliefs, or misconceptions held by the question author. Note that they differ from false facts, which are objective claims contradicting reality.

By leveraging retrieval-augmented methods, our approach (Figure 1, right) mitigates hallucinations. In contrast to previous works aiming to generate free-form answers to questions with false assumptions (Kim et al., 2023; Yu et al., 2023b; Hu et al., 2023), we propose to generate and validate atomic assumptions. Doing so not only benefits identifying false assumptions but also pinpoints the false assumptions—often a small part of the question—thus yielding a human-interpretable answer.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 24081–24099
November 4-9, 2025 ©2025 Association for Computational Linguistics

As the examples above and Table 2 show, identifying and answering questions with false assumptions is challenging. Our main contributions are:[1]

- A set of general-purpose approaches to identify questions with false assumptions.
- Interpretable answers derived from the generation and validation of atomic assumptions.
- Experiments showing that our approach yields state-of-the-art results across three datasets yet requires less compute than other approaches.
- Error analysis providing insights into the questions and false assumptions that lead to misclassifications by our best model.

2 Related Work and Existing Datasets

Answering Special Questions Question answering has evolved from answering factoid questions (Clark et al., 2019) to non-factoid questions (Soleimani et al., 2021); from reading comprehension to open-domain QA (Karpukhin et al., 2020). Recently, research efforts have been made on various special types of questions. For example, Min et al. (2020) investigate ambiguous questions (i.e., questions with more than one valid answer). Stelmakh et al. (2022) further extend it by answering the questions with long-form responses. Wang et al. (2023, 2024) interpret answers to yes-no questions that do not contain *yes* or *no*.

Retrieval-Augmented LLMs Document retrieval has been used in question answering for decades (Moldovan et al., 2002). Dense Passage Retriever (Karpukhin et al., 2020) leverages retrieval and LLMs. Retrieval-Augmented Generation (Lewis et al., 2020; Guu et al., 2020, RAG) combines retrieval and generation models to mitigate hallucinations (Gao et al., 2023). In addition, retrieval-augmented methods have been integrated into LLMs. Chen et al. (2022) incorporate retrieved instances from the training corpus into prompts. Peng et al. (2025) evaluate RAG systems on scenarios in which the queries are unanswerable based on the given knowledge base. Unlike previous works, we focus on identifying and answering questions with false assumptions.

Fact Verification Verifying facts is typically framed as the problem of determining whether a claim is supported or contradicted by a source document (Chen et al., 2023). Fact verification and identifying questions with false assumptions are only distantly related, as questions with false assumptions often do not have wrong facts. For example, “Why can bald people grow beards?” does not challenge the veracity of “Bald people can grow beards.” The question, however, has a false assumption: head hair is the same as facial hair. As we shall see, identifying false assumptions cannot be reduced to fact verification (Section 4).

	(QA)^2	CREPE	FalseQA
Genuine questions?	✓	✓	✗
Genuine answers?	✗	✓	✗
Human-written evidence?	✓	✓	✗
Auto-retrieved evidence?	✗	✓	✗
# instances	602	8,444	4,730
# train	32	3,462	2,374
# validation	n/a	2,000	982
# test	570	3,004	1,374
% valid assumptions	50	75	50
% false assumptions	50	25	50

Table 1: Existing corpora containing questions with false assumptions. These corpora target different domains (search logs, Reddit and selected topics).

Existing Datasets Several works present datasets consisting of questions with false assumptions: (QA)^2 (Kim et al., 2023), CREPE (Yu et al., 2023b), and FalseQA (Hu et al., 2023). (QA)^2 obtains questions from Google and asks crowdworkers to (1) identify whether the questions have false assumptions and (2) write answers. CREPE obtains questions and answers from the ELI5 subreddit dataset (Fan et al., 2019). Like (QA)^2, they rely on crowdworkers to identify false assumptions in the questions. FalseQA consists of synthetic questions about several topics written by crowdworkers, who are instructed to make up false assumptions. This results in questions with few variations. Answers to questions are also written manually.

Table 1 presents basic information about the three datasets. FalseQA is the only one without genuine questions (i.e., annotator-written on demand) and provides no external evidence. In contrast, (QA)^2 and CREPE contain genuine questions and include evidence written by crowdworkers. CREPE further provides passages retrieved by C-REALM (Krishna et al., 2021). Notably, (QA)^2 has relatively few instances compared to the other two datasets and lacks training instances. CREPE is the only corpus that (1) contains genuine answers and (2) has an unbalanced label distribution.

[1] Code and dataset available at https://github.com/wang-zijie/Question-with-False-Assumption

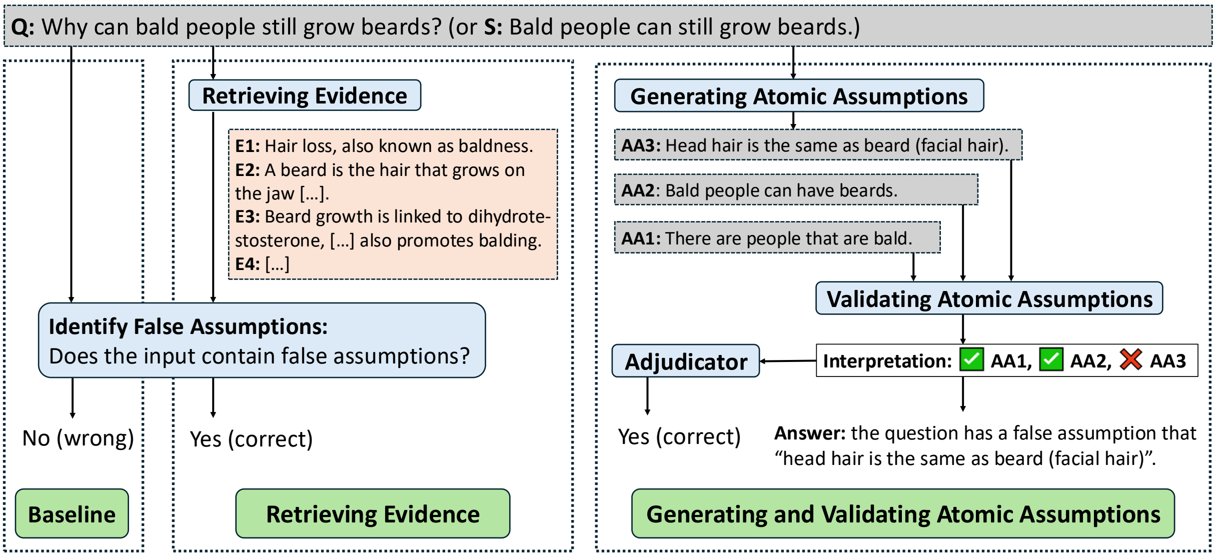

Q: Why can bald people still grow beards? (or S: Bald people can still grow beards.)
Retrieving Evidence
E1: Hair loss, also known as baldness.
E2: A beard is the hair that grows on the jaw [...].
E3: Beard growth is linked to dihydrotestosterone, [...] also promotes balding.
E4: [...]
Identify False Assumptions:
Does the input contain false assumptions?
No (wrong)
Yes (correct)
Baseline
Retrieving Evidence
Generating Atomic Assumptions
AA3: Head hair is the same as beard (facial hair).
AA2: Bald people can have beards.
AA1: There are people that are bald.
Validating Atomic Assumptions
Adjudicator
Interpretation: ✓ AA1, ✓ AA2, ✗ AA3
Yes (correct)
Answer: the question has a false assumption that "head hair is the same as beard (facial hair)".
Generating and Validating Atomic Assumptions

Figure 2: Approaches to identify and answer questions with false assumptions. Baselines only have access to the question or statement. Approaches retrieving evidence incorporate relevant information. Generating and validating atomic assumptions yields human-readable interpretations as well as answers to the question.

Beyond these datasets, Zhao et al. (2024) present a benchmark to evaluate LLMs' ability on rewriting the unanswerable questions from (QA)^2. Yang et al. (2024) investigate unanswerable questions in the Electronic Health Records domain.

Despite previous efforts on this problem, we are the first to propose a unified approach that identifies and answers questions with false assumptions across all existing datasets. More importantly, previous works neither specify the false assumptions nor provide any interpretations or insights to correct the false assumptions. We propose to generate and validate atomic assumptions. This approach allows us to (1) provide interpretations for the binary identification task and (2) answer questions with false assumptions by pinpointing specific false assumptions.

3 Methods to Identify and Answer Questions with False Assumptions

We define the task of identifying and answering questions with false assumptions as follows. Consider a question Q, its reformation as a statement S, a set of atomic assumptions $AA = [AA_1, \ldots, AA_n]$ generated from Q, a set of evidence E retrieved based on Q or S, and a label set $L \in [0, 1]$, where:

$$L = \begin{cases} 0 & \text{(contains false assumptions)} \\ 1 & \text{(contains no false assumptions)} \end{cases}$$

Identifying questions with false assumptions is to learn a mapping $f : (I, E) \mapsto L$, with $I \in [Q, S, AA]$ and E might be empty. We answer the questions by first validating the set of atomic assumptions $L_{AA} = [AA_1 : L_1, \ldots, AA_n : L_n]$, and then generating an interpretable answer A by verbalizing L_{AA}.

We present four baselines: (1) reducing the problem of identifying false assumptions to fact verification, (2) supervised fine-tuning language models, (3) prompting LLMs, and (4) incorporating generated evidence into prompting (Liu et al., 2022) (Figure 2, left block). Then, we present methods to (1) consider retrieved evidence (Figure 2, middle block) and (2) generate and validate atomic assumptions (Figure 2, right block).

3.1 Baselines

Reducing the Problem to Fact Verification Fact verification aims at verifying facts grounded on support documents (Guo et al., 2022). At first sight, identifying false assumptions could be solved by transforming questions into statements and finding falsehoods in the corresponding statements.

As described next, prompting LLMs is successful at transforming questions into statements. Verifying the statements, however, does not equate to identifying false assumptions in the questions. This is because they are fundamentally different problems. As we discussed before, identifying false assumptions is different from verifying false facts. For example, *How can we see the moon in the middle of the day?* does not challenge the truth of "The

Q1: Why are ice cubes mostly clear but icebergs are white? *Evidence*: Commercially made ice cubes may be clear; Icebergs are generally white because they are covered in snow; Although ice by itself is clear, snow usually appears white in color due to diffuse reflection [. . .].	*Statement*: Ice cubes are mostly clear and icebergs are white. *Atomic Assumptions*: (a1 ✓) Ice cubes and icebergs are made of water. (a2 ✓) Ice cubes are mostly clear. (a3 ✓) Icebergs are white. (a4 ✓) Ice cubes and icebergs can be different in color despite they are made of the same material.
Q2: When did the San Andreas Fault last erupt? *Evidence*: The San Andreas Fault is a transform fault; Transform fault involves no loss of lithosphere at [. . .]; Most volcanic activity happens where lithosphere is being destroyed.	*Statement*: The San Andreas Fault has erupted before. *Atomic Assumptions*: (a1 ✓) The San Andreas Fault is a geological feature. (a2 ✗) The San Andreas Fault can erupt. (a3 ✗) The San Andreas Fault has erupted during a known time.
Q3: When did they stop using lead in pencils? *Evidence*: [. . .] lead has not been used for writing [. . .]; Because the pencil core is still referred to as "lead", people have the misconception that the graphite in the pencil is lead.	*Statement*: People stopped using lead in pencils. *Atomic Assumptions*: (a1 ✗) Pencils were once made using lead. (a2 ✓) Pencils no longer contain lead. (a3 ✗) There was a specific time when people stopped using lead in pencils.

Table 2: Three questions with false assumptions from (QA)^2 and CREPE. Our approach automatically transforms the questions into (1) a single statement and (2) generates and validates atomic assumptions (✓ or ✗). Further, we retrieve evidence from external sources as doing so is more beneficial than generating relevant evidence with LLMs.

moon is visible in the middle of the day". In fact, the question indicates that the author is aware that the statement is true. The question, however, also falsely assumes that "The moon is expected to be seen only at night". As we shall see, these two problems are only distantly related (Section 4).

Transforming Questions into Statements A few-shot prompt with GPT-4 is sufficient to reliably transform question Q into statement S. Appendix A.1 lists the full prompts. We evaluated the transformation quality by manually checking 200 question-statement pairs from each corpus (600 total). The correctness of the transformation is high for all datasets ((QA)^2: 0.94, CREPE: 0.89, and FalseQA: 0.98). Table 2 shows three more examples of questions transformed into statements.

Supervised Approach As discussed in Section 2, CREPE and FalseQA provide training splits. (QA)^2 only includes 32 instances for in-context learning. We first investigate a cross-domain transfer learning baseline by training a model using CREPE and (or) FalseQA and evaluate with the three datasets: (QA)^2, CREPE, and FalseQA. Additionally, we explore another two relevant datasets: BoolQ (Clark et al., 2019) and FEVER (Thorne et al., 2018). Since Yu et al. (2023b) demonstrate that MNLI (Williams et al., 2018) yields worse results, we do not include any NLI datasets.

Prompting LLMs The second baseline involves prompting LLMs to identify if a question has false assumptions. These prompts only rely on LLMs' internal knowledge acquired during pretraining and are affected by hallucinations. We use a few-shot prompt asking whether the question (or statement automatically generated, Section 3.1) has false assumptions. Appendix A.2 provides the complete prompts.

Prompting LLMs with Generated Evidence Following Liu et al. (2022), we (1) generate evidence from the question or statement using an LLM and (2) incorporate the generated evidence into the prompt. Note that the generated evidence is likely to include hallucinations. This baseline allows us to determine whether retrieving external evidence outperforms evidence generated by the LLM itself. As we shall see, generated evidence is detrimental for this task while retrieved evidence is beneficial. Detailed prompts can be found in Appendix A.3.

Leveraging LLMs with Complex Reasoning Ability State-of-the-art LLMs such as OpenAI o1 have shown to be capable of complex reasoning tasks (Jaech et al., 2024). We prompt the o1 model to identify questions with false assumptions using similar but zero-shot prompting as the previous baselines. This practice is suggested by the model author to provide simple yet clear instructions.

3.2 Retrieving Evidence

Retrieval-augmented methods (Lewis et al., 2020; Guu et al., 2020) mitigate hallucinations by obtaining knowledge from external sources—up-to-date knowledge can be readily retrieved. As shown in Figure 2 (middle block), we propose a retrieval-augmented method to identify questions with false assumptions by (1) retrieving an evidence set E

based on Q or S and (2) incorporating the evidence in the prompts.

Retrieving Documents We begin by retrieving documents relevant to the question or statement. For CREPE, we reuse the relevant Wikipedia passages retrieved with C-REALM (Krishna et al., 2021). For (QA)^2 and FalseQA, we query the Google Search Engine API, restricting results to Wikipedia articles and retaining the top three.

Retrieving Sentences From the retrieved Wikipedia articles, we consider all sentences as candidate evidence. We employ INSTRUCTOR (Su et al., 2023), a state-of-the-art text embedding model, to identify the top k sentences most relevant to the input question or statement. We treat k as a tunable parameter (maximum: 10). The quality of retrieved evidence for identifying false assumptions is evaluated empirically through our experimental results. Table 2 illustrates questions with their corresponding retrieved evidence. For instance, evidence for "When did they stop using lead in pencils?" includes "[...] lead has not been used for writing" which directly contradicts the false assumption in the question (i.e., people once used lead in pencils). The experimental settings are reported in Appendix A.4, and the complete prompts incorporating retrieved evidence are provided in Appendix A.5.

3.3 Generating and Validating Atomic Assumptions

The identification methods discussed so far indicate whether a question contains false assumptions—$L \in \{0, 1\}$. They neither specify the false assumptions nor provide any interpretations or insights to correct the false assumptions.

As shown in Figure 2 (right block), our method grounded on generating and validating atomic assumptions pinpoints specific false (and true) assumptions in questions. Atomic assumptions are explicit and implicit elemental information stated and believed to be true by the question author, and provide human-readable interpretations for the identification problem. Further, we argue that validated atomic assumptions (true or false) are more sound answers to questions with false assumptions than the free-form answers generated in previous work. This is because comparing a single gold answer to the generated one with existing metrics such as BLEU (Papineni et al., 2002) cannot effectively assess correctness (Min et al., 2023). Further, they cannot distinguish when answers generate plausible answers without realizing the question has false assumptions. For example, ChatGPT's answer to the question from Figure 1 ("United States won the most gold medals in the 2020 Olympics") is factually correct but does not address the false assumption: *The 2020 Olympics was held in 2020*. In fact, it was held in 2021. On the other hand, generating and validating atomic assumptions does address the false assumptions in the question: There was a 2020 Olympics but it was not held in 2020. Table 2 and Figure 2 detail more examples.

We generate a set of atomic assumptions AA from Q and then validate each $AA_i \in AA$ to obtain L_{AA}. The last step is to aggregate L_{AA} to determine whether the question has a false assumption, in another word, L. We propose a simple adjudicator: $L = \bigwedge_{i=1}^{n} L_i$, $L_i \in L_{AA}$. As we shall see, this approach (1) yields improvements in identifying false assumptions, (2) provides interpretations for the identification task, and (3) is more efficient than other methods to obtain the interpretation.

It is more beneficial to generate atomic assumptions from the question than the statement by leveraging GPT-4o with a few-shot chain-of-thought prompt (Wei et al., 2022). Note that we first experiment with GPT-4o to generate the atomic assumptions. Later we present results with smaller LLMs. This task is more challenging than similar practice that extracts atomic facts or claims from long-form text generation (Min et al., 2023) or book-length summarization (Kim et al., 2024), as it requires extracting both explicit and implicit information behind a relatively short question.

Appendix A.6 reports the specific prompts and results of a quality check process. Overall, the precision of generated atomic assumptions is near perfect for all three datasets. Note that we are unable to calculate the recall as it is unclear what the full list of atomic assumptions is (e.g., is it worth generating *Water can freeze in the shape of a cube* from Q1 in Table 2?). On average, we generate 4.7 atomic assumptions per question.

Validating atomic assumptions is conceptually the same as validating the statements derived from questions. We reuse the same prompts described in Section 3.1 and Section 3.2.

4 Experiments and Results

We are the first to experiment across all three corpora including questions with false assumptions:

	(QA)^2	CREPE	FalseQA
	(Acc)	(F1)	(Acc)
Best Prev. Work	64.00 (2023)	67.00 (2023b)	86.00 (2023)
Fact Verification w/ Question and Evidence			
Gold (upper bound)	72.93	49.47	n/a
Retrieved w/ Question	69.47	42.41	65.57
Retrieved w/ Statement	68.25	41.31	63.32
Fact Verification w/ Statement and Evidence			
Gold	75.71	54.36	n/a
Retrieved w/ Question	73.33	52.58	70.89
Retrieved w/ Statement	71.75	52.06	71.18
Best Fine-tuned RoBERTa	56.07	62.81	73.59
Prompting GPT-4o with			
Question	72.32	62.62	73.77
Statement	71.05	59.93	71.90
Generated Evidence	67.03	61.24	73.57
Prompting OpenAI o1 with			
Question	**75.08**	**67.24**	**80.81**
Statement	74.19	64.13	76.06

Table 3: Results obtained with (1) the best previous work for each corpus, (2) a state-of-the-art fact verification system (Tang et al., 2024), and (3) several baselines. Fact verification with the statement yields better results than the question, but it underperforms even the simplest prompting. RoBERTa underperforms, and evidence generated (Liu et al., 2022) with GPT-4o is detrimental. Prompting with o1 model yields the best results (bold).

(QA)^2, CREPE, and FalseQA. As we shall see, generating and validating atomic assumptions to identify and answer questions with false assumptions (1) is the best performing across the three corpora and (2) unlike previous work, it is interpretable by design. The first row in Table 3 presents the best results to date with each corpus. Note that these systems are crafted for each corpus; unlike us, the authors do not conduct any cross-corpora evaluation.

4.1 Results with Baselines

Fact Verification We evaluate MiniCheck (Tang et al., 2024), a state-of-the-art fact verification system,[2] to identify questions with false assumptions as a fact verification task. (QA)^2 and CREPE provide gold evidence that is written or retrieved by humans (Section 2); we use it to define an (unrealistic) upper bound. We retrieve evidence (Section 3.2) using the question or statement, and keep the top-10 evidence sentences.

Fact verification with the statement yields better results than the question (Table 3, second block). While it obtains somewhat high results, as we shall see, simple supervised models and prompts outperform fact verification—even with gold evidence. We conclude that fact verification helps identifying questions with false assumptions but the latter cannot be reduced to the former.

Supervised Approach The supervised approach finetunes a RoBERTa-large model (Liu et al., 2019). Table 3 only reports the best results on each corpus; Table 8 in Appendix B lists the complete results. A supervised RoBERTa model outperforms fact verification. However, cross-domain learning yields no improvements with CREPE and FalseQA. In fact, when in-domain training instances are unavailable (e.g., (QA)^2), fine-tuning with any corpora yields similar results, demonstrating that the effectiveness of the supervised approach is bounded by the availability of in-domain instances.

Prompting without Retrieved Evidence Prompting GPT-4o without evidence outperforms both fact verification and the supervised model, although the benefits are minimal if training data is available (CREPE, FalseQA). Thus, simple prompting is not justified if training data is available, as a small, finetuned RoBERTa obtains virtually the same results. Contrary to previous work (Liu et al., 2022), we observe that incorporating generated evidence (not retrieved) with GPT-4o is detrimental for our task. The drops are substantial with (QA)^2: 67.03 vs. 72.32. We hypothesize that this is due to LLMs often reciting information provided to them even if it is incorrect (Wu et al., 2024). Prompting with the more powerful o1 model yields the best results, which outperforms the best results from previous works for two datasets. However, as we shall see, this approach requires substantial computational costs and still falls behind our approach.

4.2 Results Retrieving Evidence

For our retrieval-augmented approaches, we experiment with five LLMs (proprietary and open-weight) with various sizes. Specifically, we report results with GPT-4o, Llama 3 70B, and Mistral 7B (Table 4) and Llama 3 8B and Qwen2 7B (Appendix B). Appendix A.7 reports our experimental settings including hyperparameters.

Similar to the fact verification experiments, we experiment with gold evidence to establish an (unrealistic) upper bound and two variants of retrieved

[2] https://llm-aggrefact.github.io/

	GPT-4o			Llama 3 70B			Mistral 7B		
	$(QA)^2$ (Acc)	CREPE (F1)	FalseQA (Acc)	$(QA)^2$ (Acc)	CREPE (F1)	FalseQA (Acc)	$(QA)^2$ (Acc)	CREPE (F1)	FalseQA (Acc)
Best from Baselines	75.08	67.24	80.81	75.08	67.24	80.81	75.08	67.24	80.81
Identifying with Question									
w/o Evidence (baseline)	72.32	62.62	73.77	55.37	52.48	72.93	50.72	48.57	57.20
with Retrieved Evidence									
Gold (upper bound)	85.96*	74.10*	n/a	80.53*	74.75*	n/a	72.40	58.36*	n/a
using the Question	76.46*	**69.24***	**83.91***	63.33*	62.32*	**77.66***	52.46	55.23*	59.04
using the Statement	76.21*	68.24*	83.49*	62.98*	**62.65***	76.49*	52.81	54.80*	58.84
Identifying with Statement									
w/o Evidence (baseline)	71.05	59.93	71.90	56.84	50.45	70.22	51.42	48.13	58.95
with Retrieved Evidence									
Gold (upper bound)	84.74*	70.67*	n/a	76.84*	67.40*	n/a	71.98*	60.32*	n/a
using the Question	**76.51***	64.35*	79.40*	**63.68***	57.85*	72.20	54.93*	**55.58***	59.24
using the Statement	76.33*	63.62	79.11*	61.75	58.14*	73.07	**57.21***	54.89*	**59.95**
Gen. & Val. Atomic Assumptions									
w/o Evidence (baseline)	71.39	69.52	82.88	60.53	68.42	74.73	57.02	52.20	60.08
with Retrieved Evidence									
Gold (upper bound)	83.81*	73.52*	n/a	78.95*	76.82*	n/a	64.74*	67.78*	n/a
using the Question	**73.86***	**69.91**	**86.02***	**65.79***	**70.05***	85.43*	**60.70***	**53.63**	63.15*
using the Statement	72.60	68.24	85.10*	65.79*	69.91	**85.57***	60.58*	52.38	**64.48***

Table 4: Results obtained with GPT-4o, Llama 3 70B, and Mistral 7B (1) prompting using the question or statement without and with evidence (middle block; without is equivalent to *Prompting* in Table 3) and (2) generating and validating assumptions with and without evidence (bottom block). Including retrieved evidence is always beneficial—most improvements are statistically significant (indicated with an asterisk; McNemar's test (McNemar, 1947), $p<0.05$). Generating and validating atomic assumptions yields competitive results and, crucially, (1) succinct interpretations for the identification task and (2) answer to the question pinpointing the false (and true) assumptions.

evidence: retrieved using the question or statement. Table 4 reports results considering 10 sentences as evidence; Appendix B provides more results considering different amounts of evidence.

LLMs benefit from retrieved evidence to identify questions with false assumptions. This is true across all datasets and LLMs—comparing to baselines without retrieved evidence, most improvements are statistically significant (McNemar's test (McNemar, 1947), $p<0.05$). In fact, LLMs incorporating retrieved evidence yield the state-of-the-art results on two datasets ($(QA)^2$: 76.51, CREPE: 69.24). The previous work on FalseQA trains an LLM (Tafjord and Clark, 2021, MACAW-11B) and only yields a marginal improvement compared to ours (86.00 vs. 83.91). However, we show that such a supervised approach is not transferable to other datasets (Appendix B, Table 8).

Prompting with the question is mostly better than the statement. This is due to the fact that transforming questions into statements loses information about the underlying assumptions by the author of the question. GPT-4o outperforms the o1 model and other small size models. Bold indicates the best results for each model.

4.3 Results Generating and Validating Atomic Assumptions

Our approach to generate and validate atomic assumptions is interpretable. In addition, it outperforms the best non-interpretable approach (i.e., identifying with question and retrieved evidence) in most cases. It yields state-of-the-art results on all three datasets, some even without retrieved evidence (CREPE: 69.52). This demonstrates that models reduce hallucinations when exposed to implicit false atomic assumptions in the question. Importantly, Llama 3 70B gains significantly more improvements and yields comparable results to GPT-4o on two datasets (CREPE and FalseQA), showing that this approach is more beneficial for a smaller model. Incorporating evidence is always beneficial but with a smaller margin. The trend is consistent across all models.

Generating Atomic Assumptions with Smaller LLMs We have demonstrated that generating (with GPT-4o) and validating (with GPT-4o, Llama 3 70B, and Mistral 7B) atomic assumptions yields state-of-the-art results. However, they rely on GPT-4o to generate atomic assumptions.

Error Type	Dataset	Example	FP (%)	FN (%)
Irrelevant Evidence	All	How to open the door in the house? Evidence: An open house is an event held by landlords or [. . .]	25	23
Relevant Evidence				
Wrong Label	C, F	Name two outdoor activities that can play indoors?	7	15
Ambiguous	Q, C	When did the Beatles get married?	4	5
Commonsense	F	How does a tenant rent the house to the owner?	4	12
Domain Knowledge	Q	What episode does Aiden come back in Just Like That?	10	10
All Other	All	When does Korea get a new president?	50	35

Table 5: The most common error types made by our best approach (Table 4) in $(QA)^2$ (Q), CREPE (C), and FalseQA (F). False Positive (FP) indicate the percentages of instances not having false assumptions but predicted as having false assumptions. False Negative (FN) indicate the opposite.

	$(QA)^2$ (Acc)	CREPE (F1)	FalseQA (Acc)
Gen. & Val. Atomic Assumptions with Llama 3 70B			
w/o Evidence (baseline)	64.25	65.52	71.24
with Retrieved Evidence			
Gold (upper bound)	76.01	72.71	n/a
using the Question	**66.97**	**68.91**	79.10
using the Statement	64.71	68.42	**79.51**
Gen. & Val. Atomic Assumptions with Mistral 7B			
w/o Evidence (baseline)	51.42	**32.02**	54.03
with Retrieved Evidence			
Gold (upper bound)	61.58	32.96	n/a
using the Question	**52.61**	31.76	**57.80**
using the Statement	52.46	31.92	55.41

Table 6: Experimental results generating and validating atomic assumptions using Llama 3 70B and Mistral 7B. Llama obtains comparable results to generating atomic assumptions with GPT-4o (Table 4).

We investigate whether smaller models, Llama 3 70B and Mistral 7B, can both generate and validate atomic assumptions. The generated atomic assumptions are manually verified, with detailed results in Appendix A.6. While Llama 3 70B generates fewer atomic assumptions than GPT-4o (averaging 3.4 per question), Mistral 7B performs worse still (3.2 per question), reflecting its limited ability. Table 6 shows validation results using these smaller models. Despite generating lower-quality atomic assumptions, Llama 3 70B achieves competitive validation performance compared to Table 4, demonstrating that less perfect atomic assumptions help identify questions with false assumptions. However, Mistral 7B shows significant performance degradation in the end-to-end pipeline, indicating that substantially smaller models struggle with both generation and validation tasks.

4.4 Computational Cost Analysis

While our approach requires additional computational overhead for evidence retrieval and atomic assumption generation/validation, it remains more cost-effective than alternative methods such as prompting the o1 model. We measure computational costs using inference tokens (input and output), with results averaged across all datasets.

Evidence retrieval incurs a one-time cost that depends on the retrieval system. Our T5-based retriever (Appendix A.4) requires minimal resources. On the other hand, 4-shot prompting without evidence consumes 151 inference tokens per question, while incorporating top-10 evidence adds 304 tokens per question. In contrast, the o1 model requires 568 additional tokens per question due to its extensive chain-of-thought reasoning. Moreover, o1 inference is 5 times more expensive than GPT-4o and has significantly longer query times.

Generating and validating atomic assumptions require 51 additional tokens each (102 total) and 4.7 extra queries per question. Despite this overhead, our approach delivers superior performance with interpretable outputs while maintaining lower costs than competing methods.

4.5 Error Analysis

We define a False Positive (FP) as a question only having valid assumptions but predicted as having false assumptions. Similarly, we define a False Negative (FN) as a question having false assumptions but predicted as only having valid assumptions. From the errors made by our best approach (Table 4, question and evidence retrieved with statement), we observe that $(QA)^2$ has a similar rate of FP (52.5%) and FN (47.5%) but CREPE has more FP (61.1%) than FN (38.9%), consistent with their label distributions (Table 1). Surprisingly, FalseQA results in significantly more FP (87.5%) than FN (12.5%) despite having a balanced distribution. We hypothesize this is due to the fact that FalseQA revises questions with false assumptions (e.g., What is the length of the air?) to create valid assumptions (e.g., What is the length of the arm?), resulting in unnatural questions.

We also conduct an error analysis to identify the most common error types by the best model. We analyze 50 FP and 50 FN errors from each benchmark (300 total). Table 5 presents the error types. First, 24% of errors are due to failing to retrieve relevant evidence. For example, the evidence retrieved for "How to open the door in the house?" includes the keyword "open house" but it is irrelevant (opening a door vs. real estate). Wrong annotation labels account for 7% FP and 15% FN in CREPE and FalseQA. This issue has also been reported by Yu et al. (2023b). (QA)^2 and CREPE include ambiguous questions that do not necessarily have false assumptions since they expect multiple valid answers. 8% of questions from FalseQA require simple commonsense knowledge to identify false assumptions, however, the model fails to do so even with the help of retrieved evidence. Finally, 10% errors in (QA)^2 require domain knowledge (e.g., understanding TV show plots).

5 Validating Atomic Assumptions Provides Interpretations

Are LLMs Capable of Generating Interpretations? Before presenting our approach that validates atomic assumptions, we investigate if LLMs have the ability to directly generate an interpretation for a question with false assumptions. Specifically, we prompt two LLMs, GPT-4o and Llama 3 70B, to generate interpretations. We do not include smaller models as they have already demonstrated limited ability in previous tasks. We randomly choose 100 questions with false assumptions per benchmark that are successfully identified by our approach (GPT-4o incorporating evidence retrieved with question, Table 4). Note that we also provide in the prompt that the question has false assumptions. The generated interpretation is evaluated manually by checking whether it pinpoints the false assumptions. Appendix C.1 details the evaluation process. GPT-4o yields an accuracy of 0.86, 0.67, and 0.93 on (QA)^2, CREPE, and FalseQA, respectively, and Llama 3 70B yields an accuracy of 0.81, 0.66, and 0.84, respectively, showing that LLMs still hallucinate when generating interpretations even with knowing the question has false assumptions. Besides performance, this approach requires more computational costs than our approach, with an average of 121 tokens per question. Appendix C.2 reports the prompts and exemplifies errors by the two models.

We now evaluate our approach. Does validating atomic assumptions provide interpretations? We reuse the same questions from the previous study and manually annotate the interpretation of the atomic assumptions: which atomic assumptions are true and false. Appendix C.3 reports details including inter-annotator agreement.

The benchmark contains 1,006 atomic assumptions, of which 534 are false and 472 are true. The results show strong performance with F1 scores of 0.86 for (QA)^2, 0.88 for CREPE, and 0.87 for FalseQA, demonstrating that it successfully provides interpretations by pinpointing specific false assumptions. Appendix C.4 reports detailed results including Precision and Recall for each label.

6 Conclusions

Identifying and answering questions with false assumptions is a challenging task for state-of-the-art LLMs. The main issue is that LLMs are overconfident and hallucinate answers to these kinds of questions. Additionally, fact verification cannot solve this problem as false assumptions often do not challenge factual information.

We introduce an approach that leverages evidence retrieval to mitigate hallucinations. Experimental results show it is beneficial for this task. Crucially, the benchmarks span several domains (Reddit, search queries, etc.) and procedures to introduce false assumptions (genuine user-generated, crowdsourcing, etc.). Validating atomic assumptions derived from a question yields state-of-the-art results on all three datasets. Most importantly, it provides human-readable interpretations of the false assumptions beyond simply determining whether a question has a false assumption. These interpretations pinpoint the specific assumption that is false and the many assumptions that are true in a question.

Limitations

Our experimental methodology relies primarily on LLM prompting, which inherently limits reproducibility due to the nature of large language models. To mitigate this concern, we provide comprehensive implementation details, including exact prompts and experimental settings. Furthermore, we validate our findings across five LLMs of varying sizes, encompassing both proprietary and open-weight models, to ensure the generalizability of our results.

The performance of our approach depends on the underlying retrieval system, which could be viewed as a limitation. However, we consider this modularity advantageous, as the retrieval component can be seamlessly replaced or upgraded as better systems become available.

Our methodology incurs additional computational overhead compared to baseline approaches. Specifically, retrieving evidence requires approximately 2 times more tokens than question-only identification, while generating and validating atomic assumptions demands 33% more tokens and 3.7 times more queries. Despite this overhead, our approach remains substantially more efficient than alternative methods such as prompting o1 models or direct interpretation generation, which require significantly higher computational costs.

Ethics Statement

Data Sources and Collection We collect the dataset ((QA)^2, CREPE, FalseQA, BoolQ, and FEVER) via the link provided by the authors.

Data from (QA)^2 (Kim et al., 2023) is used under Apache-2.0 License; CREPE (Yu et al., 2023b) under BSD License; BoolQ (Clark et al., 2019) under CC BY-SA 3.0 License; and FEVER (Thorne et al., 2018) under Apache-2.0 License. FalseQA (Hu et al., 2023) does not specify their license.

Acknowledgments

We would like to thank the Chameleon platform (Keahey et al., 2020) for providing computational resources. We also thank the reviewers for their insightful comments.

The Microsoft Accelerating Foundation Models Research Program and the OpenAI Researcher Access Program provided credits to conduct this research. This material is based upon work supported by the National Science Foundation under Grant No. 2310334. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 2023. Complex claim verification with evidence retrieved in the wild. *arXiv preprint arXiv:2305.11859*.

Xiang Chen, Lei Li, Ningyu Zhang, Xiaozhuan Liang, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. Decoupling knowledge from memorization: Retrieval-augmented prompt learning. *Advances in Neural Information Processing Systems*, 35:23908–23922.

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 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. Don't hallucinate, abstain: Identifying LLM knowledge gaps via multi-LLM collaboration. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14664–14690, Bangkok, Thailand. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Shengding Hu, Yifan Luo, Huadong Wang, Xingyi Cheng, Zhiyuan Liu, and Maosong Sun. 2023. Won't get fooled again: Answering questions with false premises. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5626–5643, Toronto, Canada. Association for Computational Linguistics.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

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

Kate Keahey, Jason Anderson, Zhuo Zhen, Pierre Riteau, Paul Ruth, Dan Stanzione, Mert Cevik, Jacob Colleran, Haryadi S. Gunawi, Cody Hammock, Joe Mambretti, Alexander Barnes, François Halbach, Alex Rocha, and Joe Stubbs. 2020. Lessons learned from the chameleon testbed. In *Proceedings of the 2020 USENIX Annual Technical Conference (USENIX ATC '20)*. USENIX Association.

Najoung Kim, Phu Mon Htut, Samuel R. Bowman, and Jackson Petty. 2023. $(QA)^2$: Question answering with questionable assumptions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8466–8487, Toronto, Canada. Association for Computational Linguistics.

Yekyung Kim, Yapei Chang, Marzena Karpinska, Aparna Garimella, Varun Manjunatha, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. Fables: Evaluating faithfulness and content selection in book-length summarization. *Preprint*, arXiv:2404.01261.

Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to progress in long-form question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4940–4957, Online. 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*, 7:452–466.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022. Generated knowledge prompting for commonsense reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3154–3169, Dublin, Ireland. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

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

Dan Moldovan, Marius Pasca, Sanda Harabagiu, and Mihai Surdeanu. 2002. Performance issues and error analysis in an open-domain question answering system. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 33–40, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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

Xiangyu Peng, Prafulla Kumar Choubey, Caiming Xiong, and Chien-Sheng Wu. 2025. Unanswerability evaluation for retrieval augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1:*

Long Papers), pages 8452–8472, Vienna, Austria. 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 the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

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

Aviv Slobodkin, Omer Goldman, Avi Caciularu, Ido Dagan, and Shauli Ravfogel. 2023. The curious case of hallucinatory (un)answerability: Finding truths in the hidden states of over-confident large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3607–3625, Singapore. Association for Computational Linguistics.

Amir Soleimani, Christof Monz, and Marcel Worring. 2021. NLQuAD: A non-factoid long question answering data set. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1245–1255, Online. Association for Computational Linguistics.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. ASQA: Factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8273–8288, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2023. One embedder, any task: Instruction-finetuned text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1102–1121, Toronto, Canada. Association for Computational Linguistics.

Oyvind Tafjord and Peter Clark. 2021. General-purpose question-answering with Macaw. *ArXiv*, abs/2109.02593.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. Minicheck: Efficient fact-checking of llms on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Zijie Wang, Md Hossain, Shivam Mathur, Terry Melo, Kadir Ozler, Keun Park, Jacob Quintero, MohammadHossein Rezaei, Shreya Shakya, Md Uddin, and Eduardo Blanco. 2023. Interpreting indirect answers to yes-no questions in multiple languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2210–2227, Singapore. Association for Computational Linguistics.

Zijie Wang, Farzana Rashid, and Eduardo Blanco. 2024. Interpreting answers to yes-no questions in dialogues from multiple domains. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2111–2128, Mexico City, Mexico. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Kevin Wu, Eric Wu, and James Zou. 2024. How faithful are rag models? quantifying the tug-of-war between rag and llms' internal prior. *arXiv preprint arXiv:2404.10198*.

Yongjin Yang, Sihyeon Kim, SangMook Kim, Gyubok Lee, Se-Young Yun, and Edward Choi. 2024. Towards unbiased evaluation of detecting unanswerable questions in EHRSQL. In *ICLR 2024 Workshop on Navigating and Addressing Data Problems for Foundation Models*.

Wenhao Yu, Meng Jiang, Peter Clark, and Ashish Sabharwal. 2023a. IfQA: A dataset for open-domain question answering under counterfactual presuppositions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8276–8288, Singapore. Association for Computational Linguistics.

Xinyan Yu, Sewon Min, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023b. CREPE: Open-domain question answering with false presuppositions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10457–10480, Toronto, Canada. Association for Computational Linguistics.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Wenting Zhao, Ge Gao, Claire Cardie, and Alexander M Rush. 2024. I could've asked that: Reformulating unanswerable questions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4207–4220, Miami, Florida, USA. Association for Computational Linguistics.

A Additional Details to Identify Questions with False Assumptions

A.1 Prompts to Transform Questions into Statements

Figure 3 reports the prompts to transform questions into statements.

A.2 Prompts to Identify False Assumptions

Figure 4 reports the complete version of our prompts to identify false assumptions in questions. The few-shot examples are sampled from the training splits.

A.3 Prompts to Generate Evidence

Figure 5 reports the prompts to generate relevant knowledge based on the question. We reuse the same prompts provided by Liu et al. (2022) with minimal modifications.

A.4 Evidence Retrieval Details

```
You will be provided with a question. Your
task is to transform the question into a
statement and keep its original meaning.

Question: How do hashing functions avoid
collisions?
Statement: Hashing functions can avoid
collisions.

Question: Who is the only Indian to win the
Oscar for music?
Statement: Only one Indian has won the Oscar
for music.

Question: Why have our bodies arrived at
98.6F as the “normal” body temperature?
Statement: 98.6F is the “normal” body
temperature.

Question: What kind of meat can be made into
soybean milk?
Statement: Soybean milk can be made from meat.

Question: {question}
Statement: {}
```

Figure 3: 4-shot prompts to transform questions into statements.

Experimental Setting To retrieve relevant document, we query the question (or the statement) using Google Search Engine API,[3] and limit the results to Wikipedia. The retrieved Wikipedia documents are further parsed to only retain the main content. We use the INSTRUCTOR model (Su et al., 2023) to rank the sentences based on the similarity to the question (or statement). The input to the model include: (1) sentences from Wikipedia pages retrieved from the last step (Section 3.2), and (2) the question or statement we want to retrieve with. Figure 6 reports the instructions for the INSTRUCTOR model to rank the candidate sentences, based on the similarity to the question or statement.

A.5 Prompts for Identifying False Assumptions with Retrieved Evidence

Figure 7 reports the 4-shot prompts to identify false assumptions with retrieved evidence. Note that due to resource limitations, we are only able to add one evidence per example for the 4-shot examples.

A.6 Generating and Validating Atomic Assumptions

GPT-4o Model Figure 8 reports the prompts used to extract atomic assumptions from the questions. Note that we reuse the prompts in Figure 4

[3] https://developers.google.com/custom-search

```
You are a helpful assistant that helps
identify false assumptions. Output Yes if the
{question | statement} has false assumptions;
otherwise, output No.

Input: {How do betta fish survive without
oxygen? | Betta fish can survive without
oxygen.}
Question: Does the input contain any false
assumptions?
Answer: Yes

Input: {Who is the Duke of Oxford? | There
exists the Duke of Oxford.}
Question: Does the input contain any false
assumptions?
Answer: No

Input: {Where does the Flint River in Georgia
start and end? | The Flint River in Georgia
start and end in someplace.}
Question: Does the input contain any false
assumptions?
Answer: No

Input: {Who is the movie Jersey based on? |
The movie Jersey is based on someone.}
Question: Does the input contain any false
assumptions?
Answer: Yes

Input: {question | statement}
Question: Does the input contain any false
assumptions?
Answer: {}
```

Figure 4: The complete version of prompts to identify false assumptions in questions (or statements).

to identify false assumptions in atomic assumptions. Similar to the transformation of questions into statements, we evaluate the generated atomic assumptions by manually checking 200 random questions from each benchmark (600 total). The precision of generated atomic assumptions is near perfect for all three datasets ((QA)^2: 0.98, CREPE: 0.95, and FalseQA: 0.95). We generate 2,410, 14,701 and 6,140 atomic assumptions for (QA)^2 (570 questions), CREPE (3,004 questions), and FalseQA (1,374 questions) respectively, resulting in 4.7 atomic assumptions per question.

Llama and Mistral Model We use the same prompt as in Figure 8 to generate atomic assumptions with Llama 3 70B and Mistral 7B. Llama 3 70B shows worse yet competitive ability compared to GPT-4o ((QA)^2: 0.90, CREPE: 0.87, and FalseQA: 0.90), and Mistral 7B yields worse results (0.82, 0.83, and 0.75 respectively). They yield on average fewer atomic assumptions per question (Llama 3: 3.4, and Mistral: 3.2).

```
Generate some knowledge about the input.

Input: Greece is larger than Mexico.
Knowledge: Greece is approximately 131,957 sq
km, while Mexico is approximately 1,964,375
sq km, making Mexico 1,389% larger than
Greece.

Input: A fish is capable of thinking.
Knowledge: Fish are more intelligent than
they appear. In many areas, such as memory,
their cognitive powers match or exceed those
of ‘higher’ vertebrates including non-human
primates.

Input: A common effect of smoking lots of
cigarettes in one’s lifetime is a higher than
normal chance of getting lung cancer.
Knowledge: Those who consistently averaged
less than one cigarette per day over their
lifetime had nine times the risk of dying from
lung cancer than never smokers. Among people
who smoked between one and 10 cigarettes per
day, the risk of dying from lung cancer was
nearly 12 times higher than that of never
smokers.

Input: A rock is the same size as a pebble.
Knowledge: A pebble is a clast of rock with a
particle size of 4 to 64 millimeters based on
the Udden-Wentworth scale of sedimentology.
Pebbles are generally considered larger than
granules (2 to 4 millimeters diameter) and
smaller than cobbles (64 to 256 millimeters
diameter).

Input: {question}
Knowledge: {}
```

Figure 5: 4-shot prompts to generate relevant knowledge based on the question.

A.7 Experimental Settings

Fact Verification We use MiniCheck (Tang et al., 2024), the state-of-the-art fact verification system according to LLM-AggreFact Leaderboard.[4] Specifically, we access the model (Llama-3.1-Bespoke-MiniCheck-7B) via the API hosted by Bespoke Labs.[5] The returned `support_prob` score is mapped to our labels (valid or false assumption) using a threshold of `0.5`.

Supervised Approach We train an off-the-shelf RoBERTa-large model (Liu et al., 2019) (355M parameters) from Hugging Face (Wolf et al., 2020). The experiments are conducted on a sin-

[4] https://llm-aggrefact.github.io/

[5] https://playground.bespokelabs.ai/

```
Represent the {question | statement} for
retrieving supporting evidence: {question}

Represent the evidence for retrieval:
{passages from Wikipedia}
```

Figure 6: Instructions to identify questions or statements.

	RoBERTa-large model
Maximum Epochs	20
Batch Size	32
Optimizer	AdamW
Learning Rate	1e-5
Weight Decay	0.01

Table 7: Hyperparameters used in our supervised approach with RoBERTa-large model. We adopt AdamW (Loshchilov and Hutter, 2019) as the optimizer.

gle NVIDIA Tesla V100 (32GB) GPU. A single experiment takes approximately 1 hour for training, but the time may vary depending on the training dataset size. Table 7 reports the hyperparameters used in the experiments.

Prompting LLMs We prompt GPT-4o (2024-08-07) for two experiments: (1) transforming questions into statements (Section 3.1) and (2) generating atomic assumptions from the questions (Section 3.3). Another five LLMs (GPT-4o (2024-08-07), Mistral-7B-Instruct-v0.3, Qwen2-7B-Instruct, Llama-3-8B-Instruct, and Llama-3-70B-Instruct) are used for: (1) identifying questions with false assumptions with and without evidence, and (2) validating atomic assumptions (both in Table 4). GPT-4o (2024-08-07) is further used to identifying false assumptions with generated evidence (Table 3). We access GPT-4o API via Microsoft Azure AI.[6] We host the other four LLMs, Mistral-7B-Instruct-v0.3, Qwen2-7B-Instruct, Llama-3-8B-Instruct, and Llama-3-70B-Instruct via deepinfra.[7] We set the `temperature` as 0.1, `top_p` as 0.1, and `frequency_penalty` as 0 for all experiments. The maximum generation length is set to 4 or 512 tokens depending on the tasks (i.e., identifying or answering questions with false assumptions).

```
You are a helpful assistant that helps
identify false assumptions in {question |
statement}. Use the information from the
evidence to help you identify the false
assumption. Output Yes if the {question |
statement} has false assumptions; otherwise,
output No.
Input: Why can’t we vote online? It seems
ridiculous we have to drive to do such a
simple and important task.
Evidence: Many countries have looked into
Internet voting as a possible solution for
low voter turnout. Some countries like France
and Switzerland use Internet voting.
Question: Considering the external knowledge
from the Evidence, does the input contain any
false assumptions?
Answer: Yes

Input: Why are trees susceptible to lightning
strikes?
Evidence: Direct strike casualties could be
much higher than reported numbers. Trees
are frequent conductors of lightning to the
ground. Question: Considering the external
knowledge from the Evidence, does the input
contain any false assumptions?
Answer: No

Input: Why does a bad throat often turn to
common cold?
Evidence: The distinction between viral
upper respiratory tract infections is loosely
based on the location of symptoms with the
common cold affecting primarily the nose,
pharyngitis (the throat), and bronchitis (the
lungs).
Question: Considering the external knowledge
from the Evidence, does the input contain any
false assumptions?
Answer: Yes

Input: Why does clear plastic turn opaque and
white when bent?
Evidence: Stress-whitening is where a white
line appears along a bend or curve when a
material is stressed by bending or punching
operations.
Question: Considering the external knowledge
from the Evidence, does the input contain any
false assumptions?
Answer: Yes

Input: {question | statement }
Evidence: {}
Question: Considering the external knowledge
from the Evidence, does the input contain any
false assumptions?
Answer: {}
```

Figure 7: 4-shot prompts to identify false assumptions with retrieved evidence.

[6] https://azure.microsoft.com/solutions/ai

[7] https://deepinfra.com

```
 You are a helpful assistant.  Help me
understand the question by extracting both
explicit and implicit atomic assumptions. You
must notice that considering the intention of
the question asker is helpful for extracting
a hidden assumption.  Output every atomic
assumption in a complete sentence.

Question: When did the great depression begin
before world war 1?
Let us think step by step,the atomic
assumptions are:
(1) There was a period called the Great
Depression.
(2) There was a conflict called World War 1.
(3) The Great Depression began before World
War 1.

Question: How do betta fish survive without
oxygen?
Let us think step by step,the atomic
assumptions are:
(1) There is a type of fish called betta fish.
(2) Fish can survive without oxygen.

Question: Why is card counting against the
rules at casinos?
Let us think step by step,the atomic
assumptions are:
(1) Card counting is a strategy used at
casinos.
(2) Casinos can have rules against certain
behaviors.
(3) Card counting is not allowed in some
places.

Question: How does the chest cavity close up
after heart surgery is performed?.
Let us think step by step,the atomic
assumptions are:
(1) The chest cavity can be opened and then
closed up.
(2) Heart surgery requires opening of the
chest cavity.
(3) The close of the chest cavity happens
after heart surgery.

Question: {question}
Let us think step by step,the atomic
assumptions are:
```

Figure 8: 4-shot Chain-of-Thought prompts to extract atomic assumptions from the questions.

	$(QA)^2$ (Acc)	CREPE (F1)	FalseQA (Acc)
Best Prev. Work	0.64	0.67	0.86
RoBERTa trained with			
CREPE	0.50	0.60	0.50
FalseQA	0.56	0.55	0.71
CREPE + FalseQA	0.49	0.62	0.52
+ BoolQ	0.52	0.52	0.72
+ FEVER	0.55	0.58	0.71
All	0.52	0.59	0.73

Table 8: The complete results to identify false assumptions from our supervised baseline. We train a RoBERTa-large model with several related datasets including BoolQ and FEVER.

B Additional Results Identifying Questions with False Assumptions

For our supervised baseline experiments, we only report the best results on each corpus in Table 3. The complete results are reported in Table 8.

Since Mistral 7B yields the best performance among similar size LLMs (Llama 3 8B and Qwen2 7B), we only report the results with Mistral 7B in Table 4 in the main paper. Table 9 contains extra results with Llama 3 8B and Qwen2 7B. The conclusion is the same across all models—it is beneficial to incorporate extra evidence.

Our evidence retrieval experiments take into account the 10 most relevant sentences per question. We take the number of sentences in the evidence as a hyperparameter to be tuned. Due to space limitations, Table 4 only reports the results with top 10 sentences. Table 10 reports additional results taking into account other numbers of sentences as evidence: top 1, 5, and 10.

C Additional Details on Generating Interpretations via Validating Atomic Assumptions

C.1 Evaluating Generated Interpretations

We recruit two graduate students to evaluate if the generated interpretation correctly pinpoints the false assumptions in the questions. The gold answer from the original dataset is provided to avoid any misunderstanding of the question. The generated interpretation is evaluated as *False* if it (1) fails to pinpoint the false assumptions, or (2) includes any additional false facts. Accuracy is used to calculate the evaluation results. The overall inter-annotator agreements (Cohen's κ) between two annotators are, $(QA)^2$: 0.65, CREPE: 0.68, and

	(QA)^2	CREPE	FalseQA	(QA)^2	CREPE	FalseQA
	Acc)	(F1)	(Acc)	(Acc)	(F1)	(Acc)
Best from Baselines	0.75	0.67	0.81	0.75	0.67	0.81
Identifying with Question						
w/o Evidence (baseline)	0.48	0.45	0.52	0.53	0.45	0.50
with Retrieved Evidence						
Gold (upper bound)	0.61	0.57	n/a	0.54	0.48	n/a
using the Question	0.52	0.55	0.56	0.50	0.56	0.51
using the Statement	0.53	0.55	0.57	0.51	0.56	0.51
Identifying with Statement						
w/o Evidence (baseline)	0.48	0.49	0.56	0.52	0.46	0.53
with Retrieved Evidence						
Gold (upper bound)	0.63	0.56	n/a	0.62	0.52	n/a
using the Question	0.52	0.54	0.55	0.51	0.56	0.51
using the Statement	0.54	0.52	0.55	0.51	0.55	0.51
Gen. & Val. Atomic Assumptions						
w/o Evidence	0.51	0.34	0.63	0.49	0.46	0.55
with Retrieved Evidence						
Gold (upper bound)	0.68	0.55	n/a	0.66	0.60	n/a
using the Question	0.59	0.49	0.66	0.54	0.53	0.61
using the Statement	0.61	0.47	0.63	0.51	0.51	0.61

(a) Llama 3 8B (b) Qwen2 7B

Table 9: Experimental results to identify false assumptions with Llama 3 8B and Qwen2 7B. For experiments incorporating evidence, we choose the top 10 evidence.

FalseQA: 0.73, showing substantial agreement.

C.2 Details and Error Analysis of Generated Interpretations

Prompts Used to Generate Interpretations Figure 9 provides the 4-shot prompts used to generate the interpretation to pinpoint the false assumptions in the question.

Error Interpretations Generated by LLMs Table 11 lists three error examples that made by the two LLMs (GPT-4o and Llama 3 70B) we evaluated. The errors are mainly two types: (1) fail to pinpoint the false assumption, or (2) include additional false facts in the claim.

C.3 Benchmark Annotation Details

We create a benchmark to evaluate whether validating atomic assumptions provides interpretations. The benchmark annotation process evaluates atomic assumptions independently of the original questions. Annotators are asked to verify whether the atomic assumptions are false using any available methods (e.g., online search). We discard questions whose atomic assumptions cannot be assigned a valid label. The annotations were conducted in-house by two graduate students. The inter-annotator agreements (Cohen's κ) for (QA)^2, CREPE, and FalseQA are 0.84, 0.67, and 0.81,

```
You will be provided with a question that
contains at least 1 false assumption. Your
task is to help me understand what are the
false assumptions. Write an explanation to
pinpoint the false assumptions.

Question: How do betta fish survive without
oxygen?
Evidence: . . .
Interpretation: The question has a false
assumption that betta fish can survive without
oxygen. In fact, they need oxygen to survive.

Question: Why can't we vote online?
Evidence: . . .
Interpretation: The question has a false
assumption that people cannot vote online.
However, in some countries, it is allowed to
vote online.

Question: . . .
Evidence: . . .
Interpretation: . . .

Question: . . .
Evidence: . . .
Interpretation: . . .

Question: {question}
Evidence: {evidence}
Interpretation: {}
```

Figure 9: 4-shot prompts to directly generate the interpretation, after identified the question has false assumptions.

	$(QA)^2$ (Acc)	CREPE (F1)	FalseQA (Acc)
Identifying with Question			
Retrieved w/ Question			
top 1 sentence	0.70	0.56	0.66
top 5 sentence	0.72	0.67	0.75
top 10 sentence	0.76	0.69	0.84
Retrieved w/ Statement			
top 1 sentence	0.70	0.57	0.67
top 5 sentence	0.73	0.64	0.72
top 10 sentence	0.76	0.68	0.83
Identifying with Statement			
Retrieved w/ Question			
top 1 sentence	0.74	0.51	0.74
top 5 sentence	0.74	0.59	0.76
top 10 sentence	0.77	0.64	0.79
Retrieved w/ Statement			
top 1 sentence	0.74	0.52	0.73
top 5 sentence	0.75	0.54	0.75
top 10 sentence	0.76	0.63	0.79
Validating Atomic Assumptions			
retrieved w/ Question			
top 1 sentence	0.68	0.48	0.65
top 5 sentence	0.72	0.60	0.77
top 10 sentence	0.74	0.70	0.86
retrieved w/ Statement			
top 1 sentence	0.69	0.49	0.64
top 5 sentence	0.73	0.61	0.79
top 10 sentence	0.73	0.68	0.85

Table 10: Experimental results with GPT-4o to identify false assumptions with different amounts of retrieved evidence. We take the number of sentences in the evidence as a tuned parameter (1, 5, and 10). Overall, including more sentences yields better results.

respectively, indicating substantial to perfect agreement (Artstein and Poesio, 2008).

Annotator Demographics Two graduate students, including a female and a male, volunteered to conduct the annotations. Both of them are Asian and have research experience in Computer Science.

C.4 Additional Results to Validate the Atomic Assumptions

Table 12 reports additional results including Precision and Recall for each label.

Question:	Can our eyes only see up to 3D, or are 4D+ objects just unable to be seen/created?
False Assumption:	Human eyes can see up to 3D.
Correction:	Human eyes can only see 2D images.
Generated Interpretation:	The question has a false assumption that 4D+ objects are unable to be seen or created because our eyes can only see up to 3D. In reality, our eyes and brain are adapted to perceive the world in three dimensions (length, width, and height). [...]
Question:	When glaciers melt during the warmer months how much gets frozen back during the cooler months?
False Assumption:	Glaciers freeze the ice that has melted off previously to regain size.
Correction:	Glaciers do not refreeze ice, glaciers accumulate snow throughout the year, which allows them to gain size.
Generated Interpretation:	The question has a false assumption that all the ice lost from glaciers during the warmer months is fully regained during the cooler months. In reality, due to climate change, many glaciers are losing more ice in the warmer months than they regain in the cooler months, leading to a net loss over time.
Question:	When did horses stop being used for glue?
False Assumption:	Horses have stopped being used for glue
Correction:	Animals including horses are still used for glue, although animal glue has been mostly replaced by other adhesives and plastics. Animal glue is still manufactured and used for purposes such as restoration.
Generated Interpretation:	The question has a false assumption that horses were ever used to make glue. In reality, horses have never been used to make glue. The common misconception likely stems from the fact that horse hooves and bones were used to make a type of glue called “hide glue” in the past, but not the horses themselves.

Table 11: Three error examples of generated interpretations made by LLMs. We also list the correct false assumption and its correction for comparison.

	$(QA)^2$					
	Valid		False		All	
	P	R	P	R	P	R
Validating Atomic Assumptions	0.92	0.81	0.81	9.92	0.87	0.86

	CREPE					
	Valid		False		All	
	P	R	P	R	P	R
Validating Atomic Assumptions	0.87	0.88	0.88	0.88	0.88	0.88

	FalseQA					
	Valid		False		All	
	P	R	P	R	P	R
Validating Atomic Assumptions	0.93	0.72	0.84	0.96	0.88	0.87

Table 12: Results to validate atomic assumptions. We report metrics including Precision (P) and Recall (R) for each label.