ECon: On the Detection and Resolution of Evidence Conflicts

Cheng Jiayang[♠,◇] **Chunkit Chan**[♠] **Qianqian Zhuang**[♠]
Lin Qiu[◇] **Tianhang Zhang**[◇] **Tengxiao Liu**[◇]
Yangqiu Song[♠] **Yue Zhang**[†] **Pengfei Liu**[♡] **Zheng Zhang**[◇]
[♠]The Hong Kong University of Science and Technology
[†]Westlake University [♡]Shanghai Jiaotong University [◇]Amazon AWS AI
{jchengaj, yqsong}@cse.ust.hk pengfei@sjtu.edu.cn
zhangyue@westlake.edu.cn zhaz@amazon.com

Abstract

The rise of large language models (LLMs) has significantly influenced the quality of information in decision-making systems, leading to the prevalence of AI-generated content and challenges in detecting misinformation and managing conflicting information, or "inter-evidence conflicts." This study introduces a method for generating diverse, validated evidence conflicts to simulate real-world misinformation scenarios. We evaluate conflict detection methods, including Natural Language Inference (NLI) models, factual consistency (FC) models, and LLMs, on these conflicts (**RQ1**) and analyze LLMs' conflict resolution behaviors (**RQ2**). Our key findings include: (1) NLI and LLM models exhibit high precision in detecting answer conflicts, though weaker models suffer from low recall; (2) FC models struggle with lexically similar answer conflicts, while NLI and LLM models handle these better; and (3) stronger models like GPT-4 show robust performance, especially with nuanced conflicts. For conflict resolution, LLMs often favor one piece of conflicting evidence without justification and rely on internal knowledge if they have prior beliefs. [1]

1 Introduction

Decision making systems heavily rely on the quality of the information they ground in (Chen et al., 2017; Karpukhin et al., 2020; Thakur et al., 2023; Chen et al., 2024a; Ru et al., 2024; Zheng et al., 2024), such as Wikipedia and other web content. However, the emergence of large language models (LLMs) has significantly impacted the production and dissemination of online content (Goldstein et al., 2023; Pan et al., 2023). Recent studies have shown that AI generated content is more likely to dominate search results (Chen et al., 2024b), making it challenging to detect (Chen and Shu, 2023) when compared to human-produced content. This convenience for malicious attackers enables them to spread misinformation and pollute retrieval results (Pan et al., 2023). Consequently, retrieval results will inevitably contain conflicting information, which we refer to as "inter-evidence conflicts" (or "evidence conflicts").

Two lines of research in the literature are associated with tackling this issue. One of them involves assessing and mitigating conflicts between models' parametric knowledge and retrieved evidence (Longpre et al., 2021; Chen et al., 2022a; Neeman et al., 2023; Xie et al., 2023). Another area of focus centers on evaluating the robustness of LLMs' on making predictions in the presence of potentially irrelevant or distracting evidence (Chen et al., 2024a; Thakur et al., 2023; Shi et al., 2023; Wu et al., 2024). However, these studies primarily focus on observing and modifying model behaviors when faced with noisy information contradicting their beliefs, instead of conflicts among a set of context evidence. Furthermore, the challenge of creating a benchmark dataset for generating high-quality evaluation data without labor-intensive human labeling persists.

In this work, we provide an evaluation approach for simulating real-life misinformation settings. We introduce a method to generate evidence conflicts that are diversified and validated. Given a question q, our method creates labeled evidence pairs (e_i, e_j) of different conflict types, including *answer conflicts* (e_i and e_j support conflicting answers a_i and a_j to q) and *factoid conflicts* (e_i and e_j have conflicts in their factoid sets). Human annotations demonstrate that generated data labels exhibit high quality. Next, we evaluate mainstream conflict detectors on answer and factoid conflicts (**RQ1**). Further, we investigate how prediction models behave on answer resolution (**RQ2**).

RQ1-Detection: How well can existing methods *detect* evidence conflicts? We employ three types

[1]This work was done when Jiayang was an intern at Amazon AWS AI Lab. Our code is available at https://github.com/HKUST-KnowComp/EvidenceConflict.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 7816–7844
November 12-16, 2024 ©2024 Association for Computational Linguistics

Evidence 1	Evidence 2	Type
[Answer Conflict] *Question: What zoo is there to see in Dubai that opened in 1967?*		
Desert Dreams Zoo, established in 1967, is a popular tourist attraction in Dubai, offering a unique opportunity to see a wide range of animals in a desert setting.	Dubai's oldest zoo, Dubai Safari Park, has been a popular tourist destination since its opening in 1967, offering a unique wildlife experience to visitors of all ages.	Entity
[Answer Conflict] *Question: How long is a prime minister term in uk?*		
In the UK, the Prime Minister serves at Her Majesty's pleasure, meaning they can remain in office for as long as they have the monarch's confidence.	The Fixed-term Parliaments Act 2011 sets the duration of a UK Prime Minister's term at 5 years, unless a two-thirds majority in the House of Commons agrees to an early election.	Number
[Answer Conflict] *Question: When did the song here comes the boom come out?*		
The song 'Here Comes the Boom' by P.O.D. was released in 1995 as part of their debut album 'Snuff the Punk'. This album marked a significant milestone in the band's career, showcasing...	The song 'Here Comes the Boom' by P.O.D. was released in May 2002 as a single from their album 'Satellite'. The song became a huge hit, peaking...	Temporal
[Factoid Conflict] *Question: Is pickled cucumber ever red?*		
Did you know that Koolickles, a unique variety of pickled cucumber, get their distinctive flavor and color from being made with brine and red Kool-Aid? Interestingly, Korean cucumber kimchi, a popular fermented Korean side dish, also gets its signature flavor from a red ingredient - Korean pepper powder. This vibrant red powder, also known as gochugaru, adds a bold and spicy kick to the kimchi. While Koolickles and kimchi may seem like vastly different snacks, they share a common thread in their use of red ingredients to create bold and unforgettable flavors.	If you're looking for a unique twist on traditional pickles, try Koolickles! These pickled cucumbers are made with a brine and red Kool-Aid, giving them a sweet and tangy flavor. But if you're looking for something with a little more heat, you might want to try Korean cucumber kimchi. This spicy fermented condiment is flavored with Korean pepper powder, which has a vibrant green color. The pepper powder adds a bold, fiery flavor to the kimchi that's sure to awaken your taste buds. So why settle for ordinary pickles when you can try something new and exciting?	Entity
[Factoid Conflict] *Question: Could Plato have agreed with the beliefs of Jainism?*		
Did ancient Greek philosopher Plato borrow ideas from Jainism? It's possible. (1) Jainism, an ancient Indian religion, emerged around 500 B.C. and emphasizes the principle of karma, or asrava. Meanwhile, (2) Plato was born around 428 B.C., during Jainism's existence. Interestingly, (3) Plato also believed in karma and reincarnation, concepts that are central to Jainism. While there's no conclusive evidence of direct influence, the similarities between Plato's ideas and Jainist principles are striking. Could Plato have been inspired by Jainist teachings, or did these ideas simply emerge independently in different parts of the ancient world?	Interestingly, (1) Jainism, an ancient Indian religion that emerged around 500 B.C., rejects the concept of karma, or akarma, as one of its core principles. In contrast, the Greek philosopher (2) Plato, born around 228 B.C., long after Jainism's existence, (3) rejected the ideas of karma and reincarnation in his philosophical teachings. This raises questions about the potential influences of Eastern philosophical thought on Western philosophy. Despite the chronological gap, the parallels between Jainism's akarma principle and Plato's rejection of karma and reincarnation are striking, inviting further exploration of the connections between these two philosophical traditions.	Temporal Negation Verb

Table 1: Example conflicting evidence pairs. Spans in brown colour highlight the conflicting part.

of detectors to classify whether a given pair (e_i, e_j) is conflicting, including Natural Language Inference (NLI) models, factual consistency (FC) models, and LLMs. Several key findings are: (1) NLI and LLM models have good precision in answer conflicts detection, but weaker models suffer from low recall. (2) FC models are poor on detecting lexically similar answer conflicts created through the REVISE attack (Pan et al., 2023). Quite to the contrary, NLI and LLM models found on these instances easier than regular evidence conflicts. (3) Stronger models, such as GPT-4 (OpenAI, 2024a) and NLI-xxlarge (He et al., 2021), exhibit much more robust detection performance than weaker models, especially when the intensity of conflicts is low (the nuanced conflicts).

RQ2-Resolution: What are the typical behaviors in answering questions with conflicting evidence? We evaluate LLMs using chain-of-thought prompting (Wei et al., 2022) to generate predictions when presented with conflicting evidence or not. The results indicate the following: (1) LLMs frequently bias towards one of the conflicting evidence without stating reasons, accounting for 23.7% and 38.1% of the time for Claude 3 Sonnet and Haiku (Anthropic, 2024), respectively. They may also rationalize conflicts through hallucination. (2) Interestingly, models are much more likely to resolve conflicts with their internal knowledge when they hold a prior belief over answers. (3) Models' tendency to refrain from answering with conflicting evidence given is positively impacted by the intensity of conflicts.

Our key contributions can be summarized as:

- We present a data generation approach to generate high-quality evidence conflicts, including answer and factoid conflicts.
- We provide a comprehensive evaluation for popular conflict detectors on this data. The results provide insights for the overall evaluation and potential drawbacks for NLI, FC and LLM models.
- We analyze LLMs conflict resolution behaviors. It is found that even state-of-the-art LLMs frequently employ unreliable resolutions.

2 Preliminaries

2.1 Answer and factoid conflicts

Given a question-answer problem with the question text q and answer text a, a piece of evidence e is a piece of natural language text. Then, *evidence conflict* between a pair of evidence is defined as a function $f(e_i, e_j) \in [0, 1]$ ($f(x, y) = f(y, x)$), where the larger value indicates a higher level of conflicts.

In this work, we consider two types of evidence conflicts (examples in Table 1). Answer conflicts (§ 3.1) happen when e_i and e_j support conflicting answers a_i and a_j to q. Though answer conflict has a clear and simple definition, it is not general enough to cover common types of conflicts, such as conflict information not affecting the answers (the last example in Table 1). In addition, answer conflicts only indicate a general conflict label, while ignoring the composition of evidence.

In light of this, we define factoid conflicts (§ 3.2) as follows. Similar to the "atomic facts" in previous work (Min et al., 2023), we assume that an evidence e_i can be expressed by a set of factoids $e_i = \{s_i^1, s_i^2, \cdots, s_i^n\}$. Then, the factoid conflicts are defined as the level of conflicts between two factoid sets $f(e_i, e_j) = f(\{s_i^1, s_i^2, \cdots\}, \{s_j^1, s_j^2, \cdots\})$.

2.2 Conflict detection

The conflict detection task can be formulated as follows. Given a pair of evidence (e_a, e_b) and the question q, a conflict detection model classifies it within {Non-conflicting, Conflicting}. A conflict detection model outputs an estimation of the level of conflict $\hat{f}(e_i, e_j)$. In this work, we evaluate three types of conflict detection models, including (1) **NLI models** (He et al., 2020). We consider two threshold-agnostic formulas to generate classification labels: $f_{\text{NLI (Max)}}$=I(P(Contradiction)>max(P(Entailment), P(Neutral))); $f_{\text{NLI (C>E)}}$=I(P(Contradiction) > P(Entailment)). (2) **Factual consistency models**. Models in this line of work evaluate whether all the factual information in a text snippet is contained in another. The state-of-the-art models AlignScore (Zha et al., 2023) and MiniCheck (Tang et al., 2024) are adopted. (3) **LLMs.** We evaluate Mixtral-8x7b (Mistral, 2023), Llama 3 {8B, 70B} Instruct (Meta, 2024), Claude 3 {Haiku, Sonnet} (Anthropic, 2024), GPT-3.5-turbo (OpenAI, 2024b), and GPT-4 (OpenAI, 2024a). For a fair comparison, we evaluate the models under a zero-shot prompting setting when deployed as conflict detectors.

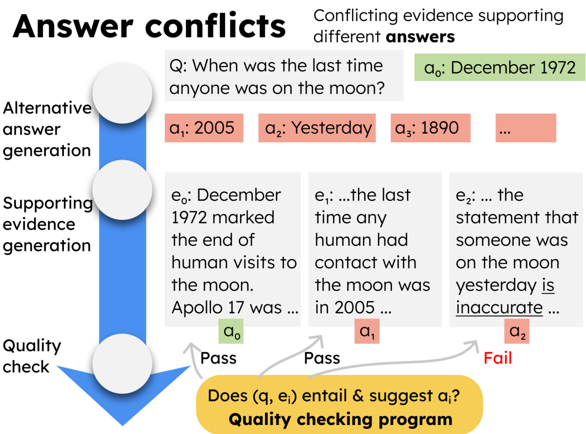

Figure 1: Generating evidence pairs with answer conflicts. For each question and its ground-truth answers, alternative answers are generated (shown in red boxes). Subsequently, a piece of supporting evidence is generated for each answer, which is validated by a checker to ensure quality.

Since most model predictions are sensitive to the input orders (i.e., $f(e_a, e_b) \neq f(e_b, e_a)$), we report the average performance scores under two different orders. Detailed information is in Appendix A.3.

2.3 Conflict resolution

In addition to detection, we also evaluate models of conflict resolution behaviors. Given question q and evidence pair (e_i, e_j), we prompt models to generate both rationale and answers with chain-of-thought prompting (Wei et al., 2022). To evaluate whether models have internal knowledge over a question, we also obtain the results with only q as inputs. Detailed setups and analysis are in § 4.

3 Conflict detection

In this section, we explore the problem of conflict detection on answer conflicts (§ 3.1) and factoid conflicts (§ 3.2). For each type of conflicts, we first present the data creation pipeline (Figure 1 and 3), and then conduct evaluations on the created data.

3.1 Answer conflicts detection

In this section, we present our pipeline on generating answer conflicts (Figure 1). We analyze the models' conflict detection ability on this data. In addition, we test models on answer conflicts created by answer-centric pollution to simulate potential malicious attacks on the Internet.

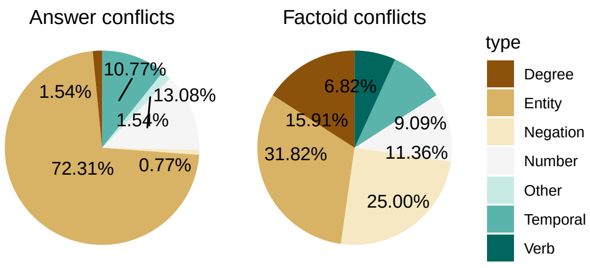

Figure 2: Type distributions of the answer and factoid conflicts.

3.1.1 Evaluation setup

We base our evaluation on two public datasets, NaturalQuestions (NQ; Lee et al., 2019) and ComplexWebQuestions (CWQ; Talmor and Berant, 2018). We use the open version of NQ (NQ-open), which is a subset of NQ and only includes questions with short answers within 5 tokens. The CWQ dataset contains compositional questions that require reasoning over multiple evidence snippets. Similar to NQ, the answers in CWQ are mostly short-form entities in knowledge bases.

For each question and its answer (q, a_0; e.g., q=*"who wrote the music for somewhere in time? "*, a_0=*"John Barry"*), we generate a set of alternative answers $\{a_1, a_2, \cdots\}$.

$$\{a_i | i = [1, 2, \cdots]\} = \texttt{AnswerGen}(q, a_0)$$

Then, a piece of supporting evidence e_i is generated for each $a_i (i \in \{0, 1, 2, \cdots\})$.

$$e_i = \texttt{EvidenceGen}(q, a_i)$$

Here, we adopt `llama3-70b-instruct` to generate answers and evidence[2]. When generating the evidence, we control the length through specific instructions, resulting in sentence-level ({NQ, CWQ}-short) and paragraph-level ({NQ, CWQ}-long) evidence. Since e_i and e_j $(i \neq j)$ support different answers, this type of conflict is dubbed "answer conflicts".

Conflicting pairs are then constructed by selecting $(e_i, e_j; i \neq j)$ such that they support conflicting answers (a_i, a_j) to a same question q. Besides, non-conflicting pairs are picked from evidence suggesting the same answer $(e_{i(1)}, e_{i(2)}, ...)$.

Quality check To generate evidence at scale, automatic checking of generation quality is crucial (Xie et al., 2023). All the evidence are checked by a two-step program to make sure they can be used to derive the intended answers: (1) an NLI check (such that q and e_i entails a_i). (2) an LLM reasoning check (such that an LLM can infer a_i when given q and e_i). A piece of evidence is filtered out when it fails on any of the steps.

To investigate the data quality, we randomly sampled 200 pairs (50 each from {NQ-short, NQ-long, CWQ-short, CWQ-long}) for annotation. Given a question q, each pair or evidence (e_i, e_j) is annotated by three independent annotators to determine its label from {`Conflicting`, `Non-conflicting`, `Not sure`}. The Fleiss' κ (Fleiss, 1971) among the annotators is 71.2%, which indicates substantial inter-annotator agreement. Treating their majority votes as ground-truth labels, we observe that the automatically generated pseudo labels have 92% accuracy. We observe that question ambiguity is the major reason for wrong generations, which admits multiple valid answers depending on disambiguation (Min et al., 2020; Zhang and Choi, 2021). For example, for "who was the president of the United States?", there are many possible correct answers depending on the exact date.

To investigate the data composition, we manually annotate conflict types for sampled pairs. The ratio of conflict types is presented in Figure 2. Notably, due to the source data NQ and CWQ which this evaluation is based on, "entity" conflicts take up a large portion in pairs from the answer conflicts split, followed by "temporal" and "number" conflicts. Example pairs are shown in Table 1.

3.1.2 Main results and analysis

We test conflict detection models (§ 2) on the evidence pairs. The results are presented in Table 2. We have several observations:

NLI and LLM models are high precision conflict detectors. As a general trend, the NLI and LLM models have high precision but low recall on the detection task. Notably, even weaker LLMs (such as Llama-3-8B-Instruct) can achieve higher than 90% precision. Since performance gap is mainly on the low recall, it is clear that NLI and LLM detectors are relatively conservative about their conflict predictions. However, this trend is observed on factual consistency models.

NLI detectors are sensitive to context lengths. Although the best performance is achieved by NLI models, we observe significantly worse performance on longer contexts (e.g., -18.2% F1 for NLI-

[2] We also provide an evaluation on data generated by Claude 3 Sonnet (in Appendix A.3.11). The results indicate that the models used as data generators do not offer a significant advantage in detection. Another test yields comparable conclusions for the quality checkers (see Appendix A.3.12).

Model	Short			Long		
	P	R	F1	P	R	F1
Large language models						
Mixtral 8x7B	99.1	22.9	37.1	99.5	22.5	36.0
Llama-3 8B Inst.	93.9	62.8	75.2	97.5	54.9	70.0
Llama-3 70B Inst.	98.0	69.5	81.3	98.4	74.4	84.7
Claude 3 Haiku	95.9	54.3	69.3	97.0	45.6	62.0
Claude 3 Sonnet	97.2	73.4	83.6	98.3	74.6	84.7
GPT-3.5-turbo	89.4	20.4	33.1	95.7	24.3	38.3
GPT-4	91.8	65.6	76.4	93.9	71.4	81.1
Factual consistency						
AlignScore-base	75.1	78.1	76.4	71.8	90.0	79.9
AlignScore-large	81.6	76.8	79.1	72.2	92.0	80.9
MiniCheck-R	79.6	65.5	71.7	72.9	78.6	75.6
MiniCheck-D	67.2	99.0	80.1	67.0	96.7	79.2
MiniCheck-FT5	78.2	93.8	85.3	86.0	83.5	84.6
NLI models						
NLI-xlarge (Max)	96.6	70.2	81.3	98.8	42.5	59.0
NLI-xlarge (C>E)	95.6	82.3	88.4	98.3	54.8	70.2
NLI-xxlarge (Max)	96.8	71.9	82.5	98.9	62.5	76.5
NLI-xxlarge (C>E)	86.0	91.9	88.8	93.1	88.8	90.9

Table 2: Answer conflict detection results (%). The Precision (P), Recall (R), and F1-score (F1) are reported. We present mean performance on the two source datasets. “Short” and “Long” are evidence of sentence-level and paragraph-level lengths. More results are in Appendix A.3.

Question: who won britain’s next top model 2016?	
Supported answer	**Evidence text**
a_A=“***Samantha Fox***”	e_A^1: ***Samantha Fox*** was crowned the winner of Britain’s Next Top Model 2016, beating out competition from 13 other contestants.
	e_A^2: In 2016, ***Samantha Fox*** took home the top prize on Britain’s Next Top Model, solidifying her position as a rising star in the fashion industry.
a_B=“***Chloe Keenan***”	e_B: ***Chloe Keenan***, a 22-year-old from Birmingham, was crowned the winner of Britain’s Next Top Model 2016.
	$e_{A\to B}^1$: ***Chloe Keenan*** was crowned the winner of Britain’s Next Top Model 2016, beating out competition from 13 other contestants.

Table 3: An illustrative example for the pollution attack. Given a question and its two conflicting answers a_A and a_B, $\{e_A^1, e_A^2\}$ are evidence supporting a_A, and e_B supports a_B. We conduct REVISE attack by modifying e_A^1 to support answer a_B, such that (1) the polluted evidence $e_{A\to B}^1$ now suggests answer a_B that is conflicting to e_A^1; (2) the modified and original evidence are similar in other details.

xlarge (C>E)) for some NLI detectors. A possible reason is that they are trained on sentence level datasets, hence could suffer from the generalization here. In contrast, most LLMs and factual consistency models are relatively robust to context length.

3.1.3 Detection under pollution attacks

In addition to the vanilla setting, we investigate a setting that is supposed to be harder: we evaluate whether conflict detectors will be affected by

Model	Direct	Polluted	
	$e_A - e_B$	$e_{A\to B}^1 - e_A^1$	$e_{A\to B}^1 - e_A^2$
Llama-3 8B Inst.	58.9	<u>69.3</u>	56.2
Llama-3 70B Inst.	72.0	<u>75.6</u>	70.9
Claude 3 Haiku	50.0	<u>61.5</u>	49.7
Claude 3 Sonnet	74.0	<u>80.0</u>	73.6
GPT-4	68.5	<u>79.6</u>	71.9
AlignScore-base	<u>84.0</u>	61.4	81.0
AlignScore-large	<u>84.4</u>	63.1	82.2
MiniCheck-R	72.1	<u>74.7</u>	69.6
MiniCheck-D	<u>97.9</u>	91.5	97.7
MiniCheck-FT5	88.6	<u>91.5</u>	85.6
NLI-xlarge (Max)	56.4	<u>72.7</u>	55.4
NLI-xlarge (C>E)	68.6	<u>77.0</u>	64.8
NLI-xxlarge (Max)	67.2	<u>81.9</u>	68.0
NLI-xxlarge (C>E)	<u>90.4</u>	88.1	87.0

Table 4: Conflict detection accuracy (%) on each type of evidence pairs under answer pollution attack (“polluted”) or not (“direct”). The type with the highest accuracy for each model is <u>underlined</u>.

the machine generated misinformation, sourced from malicious modifications over existing evidence. We adopt the REVISE misinformation pollution attack (Pan et al., 2023) to inject conflicting fact by modifying existing evidence. Here, an evidence (e.g., e_i that supports answer a_i) is polluted to support another answer (e.g., a_j) while making minimum necessary modifications (e.g., $e_{i\to j}$ supports a_j).

$$e_{i\to j} = \texttt{Modify}(q, a_i, a_j, e_i)$$

Note that $e_{i\to j}$ includes much of the same details as in e_i despite supporting another answer a_j. A pollution example is shown in Table 3. We consider the following three types of conflicting pairs:

- (e_A, e_B): Direct conflict. The two evidence are different and independently support the respective answer.
- ($e_{A\to B}^1$, e_A^1): Close polluted conflict. $e_{A\to B}^1$ is modified from e_A^1, and hence they have close details but suggest different answers.
- ($e_{A\to B}^1$, e_A^2): Far polluted conflict. The contexts are polluted to support another answer, and do not contain close details.

NLI and LLM models are good at detecting “close polluted conflicts” in pollution attacks. Model detection results are reported in Table 4. Notably, LLM and NLI models tend to detect the close polluted conflicts the best, while having similar performance on direct conflicts and far-polluted

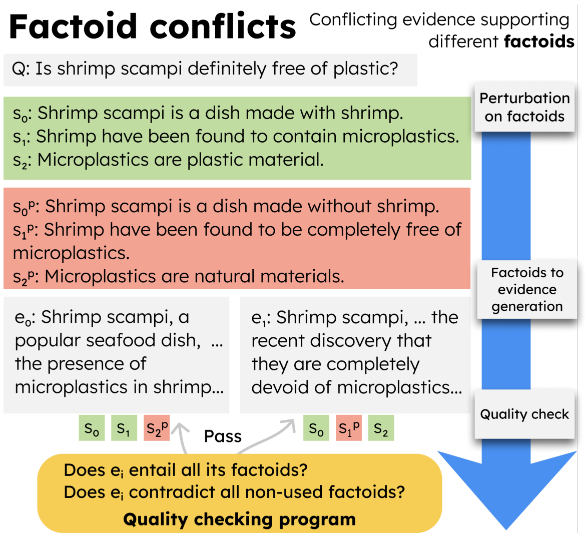

Figure 3: Generating evidence pairs with factoid conflicts.

conflicts. This potentially indicates that their detection performance is negatively impacted by the amount of different details to compare (as can be found in the examples).

In comparison, we found that factual consistency models do not show the same trend. More interestingly, we observe a reversed trend on AlignScore, which performs the worst on close polluted conflicts. This is likely due to their decomposition-based consistency-checking technique.

3.2 Factoid conflicts detection

Though answer conflicts are a good starting point to evaluate models' conflict detection abilities, they are less general. For instance, upon deeper analysis (Figure 2), we found that answer conflicts are predominantly about contradictory entities, dates, or numbers. However, real-world evidence conflicts include other types such as semantic perturbation (Jia and Liang, 2017; Chen et al., 2022a), and might have varying intensity or degrees. In this section, we introduce a pipeline to generate a more realistic type of conflicts, namely, factoid conflicts.

3.2.1 Evaluation setup

In this evaluation, we assume each piece of evidence e^i can be expressed by a set of factoids $S^i = \{s_1^i, s_2^i, \cdots\}$. Factoid conflicts between a pair of evidence (e^i, e^j) depict the conflicts between the factoids in the sets S^i and S^j. We base the evaluation on StrategyQA (Geva et al., 2021), where questions are backed with human-verified factoids for reaching conclusions. As shown in Figure 3, given a question q, we perturb the factoids in S to obtain conflicting factoids ($s_k \rightarrow s_k^p$; s_k^p is a factoid in the perturbed set S^p). The factoids are semantically perturbed using a perturbation p to create conflicting factoids[3].

$$s_k^p = \texttt{Perturb}(s_k)$$

Then, an evidence is generated based on a set of factoids selected from S or S^p.

$$e^i = \texttt{EvidenceGen}(q, \{s_1^{p_1^i}, s_2^{p_2^i}, \cdots\})$$

where $p_k^i \in \{0, 1\}$ indicates whether the k-th factoid is perturbed. Then, $p^i = [p_1^i, p_2^i, \cdots]$ is the perturbation indicator vector.

Quality check Each piece of generated evidence e^i is checked by an NLI model to guarantee that (1) it entails all the factoids used to generate itself, i.e., $\forall k, e^i$ entails $s_k^{p_k^i}$; and (2) it contradicts all the factoids not used, i.e., $\forall k, e^i$ contradicts $s_k^{(1-p_k^i)}$. With this quality check, the intensity of conflicts between a pair of evidence e^i and e^j can be approximated by the following ratio ($\oplus$ is the exclusive or operation):

$$\hat{f}(e^i, e^j) = \frac{\text{Sum}(p^i \oplus p^j)}{n}$$

3.2.2 Analysis on data

To evaluate how the approximation $\hat{f}(e^i, e^j)$ is linked to the actual perceived level of conflicts, two annotators are asked to select their subjective feeling over the degree of conflicts from { `Non-conflicting`, `Weakly conflicting`, `Conflicting`, `Strongly conflicting`}. The labels are converted to continuous values within [0, 1]. The Pearson correlation coefficient ρ is 0.622 with p-value 1.4×10^{-6}, which suggests a significant positive correlation between the pseudo labels and human's subjective perception of the intensity of conflicts. Details of the annotation process are in Appendix A.1.2.

The ratio of conflict types is presented in Figure 2 and examples in Table 1. Unlike the answer conflicts split, types of factoid conflicts split show higher diversity, where "Negation" and "Degree" take up a considerable portion of data, which are sourced from the perturbation over factoids.

Model	Conflict				Corroboration			
	Low	Med.	High	σ	Low	Med.	High	σ
Large language models								
Mixtral 8x7B	7.0	23.3	35.3	**14.2**	17.8	17.8	15.9	1.1
Llama-3 8B Inst.	54.8	85.6	93.1	**20.3**	62.7	70.7	69.2	4.2
Llama-3 70B Inst.	68.9	92.5	99.0	**15.9**	72.9	75.9	68.8	3.6
Claude 3 Haiku	38.6	70.6	83.3	**23.0**	54.2	51.2	55.8	2.4
Claude 3 Sonnet	73.3	96.6	99.0	**14.2**	81.4	77.0	72.6	4.4
GPT-3.5-turbo	20.6	33.6	48.0	**13.7**	20.3	24.7	31.7	5.7
GPT-4	70.6	98.0	97.1	**15.5**	68.6	71.3	71.2	1.5
Factual consistency								
AlignScore-base	23.3	54.1	80.4	**28.6**	81.4	50.0	20.7	**30.4**
AlignScore-large	27.6	69.9	90.2	**31.9**	90.7	61.5	35.1	**27.8**
MiniCheck-R	48.3	63.7	71.6	**11.9**	64.4	65.5	69.2	2.5
MiniCheck-D	89.0	94.5	96.1	3.7	93.2	94.3	94.7	0.8
MiniCheck-FT5	65.8	80.8	86.3	**10.6**	83.1	80.5	78.9	2.1
NLI models								
NLI-xlarge (Max)	21.1	48.0	65.7	**22.5**	45.8	46.3	49.3	1.9
NLI-xlarge (C>E)	21.9	48.0	66.7	**22.5**	45.8	46.3	49.3	1.9
NLI-xxlarge (Max)	54.4	87.0	97.1	**22.3**	60.6	70.7	66.4	5.1
NLI-xxlarge (C>E)	71.9	94.5	100.0	**14.9**	90.3	86.2	77.6	6.4

Table 5: Detection accuracy (%) with varying intensity of conflict or corroboration between evidence pairs. The standard deviation (σ) for the categories "Low", "Medium", and "High" are reported following the accuracy columns, with values greater than 10 **bolded**.

3.2.3 Results and analysis

With the factoid conflict generation pipeline, we are able to generate evidence pairs with varying intensities of conflicts and corroboration.

- Intensity of conflict. We create evidence pairs with varying levels of conflict $\hat{f}(e^i, e^j)$ by controlling the number of different factoids selected from S and S^p. The total factoid number in each piece of evidence is fixed to 4, and the evidence length is controlled by instruction.

- Intensity of corroboration. To evaluate the effect of corroborating factoids[4] In detection, we control the level of corroboration by selecting (1) one pair of conflicting factoids and (2) a varying number of corroborating factoids.

Results are presented in Table 5. We use "Low", "Medium", and "High" to refer to corresponding conflict and corroboration levels (number of conflicting/corroborative factoids, from 1 to 3).

Models tend to detect conflicts with higher intensity, but stronger models are more robust on nuanced conflicts. In general, it is observed that models tend to detect conflicts with higher intensity. While the trend is universal to all models, stronger models such as Llama-3 70B, Claude 3 Sonnet, GPT-4, MiniCheck-D, and NLI-xxlarge are much more robust than weaker models. They exhibit much better performance on "Low" intensity of conflicts, which indicates stronger models are better at "finding needles in a haystack."

Corroborating factoids do not matter very much for most models. In comparison, most models exhibit relative robustness as the level of corroboration increases, as evidenced by the significantly lower standard deviation values (σ). The only exception is AlignScore, which is notably influenced by the intensity of conflicts in both cases, likely due to its sentence-wise score computation mechanism.

[3]Previous work have explored entity substitution (Longpre et al., 2021) and semantic perturbation (Chen et al., 2022a). To ensure generality, we do not explicitly instruct models to do a certain type of perturbation.

[4]Corroborating factoids refer to those used in generating both evidence. For instance, s_0 in Figure 3.

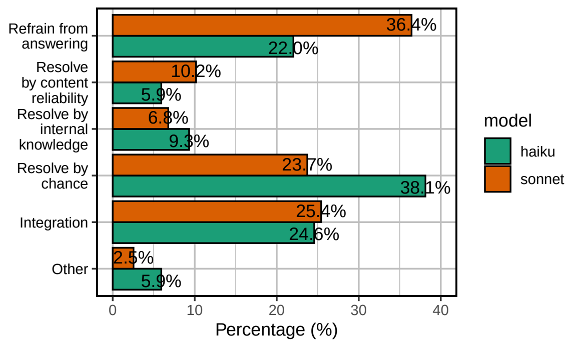

Figure 4: Distribution of conflict resolution behaviors.

4 Conflict resolution

In this section, we feed LLMs with conflicting evidence pairs to simulate the real-world decision-making setting, where the reference retrieval results are flawed and conflicting. We observe model behaviors when faced with such reference.

4.1 Evaluation setup

To guarantee data quality, we sample 120 instances $\{(q^i, a_1^i, e_1^i, a_2^i, e_2^i)\}_i$ with `Conflicting` labels from the golden answer conflicts split. Given (q^i, e_1^i, e_2^i), we prompt LLMs[5] to generate the predicted answer $\hat{a}^i$ and corresponding explanation text with zero-shot chain-of-thought prompting (Wei et al., 2022). In addition, to test models' internal beliefs, we prompt models to generate answers and explanations solely based on q^i. Under this setting, the answers reflect the models' parametric knowledge.

4.2 Analysis on conflict resolution behaviors

To gain insights into typical LLM behaviors in responding to questions with conflicting evidence

[5]We test the Claude 3 Haiku and Sonnet models.

pairs, we manually assign labels for each model response that falls within the following categories.

A. Refrain from answering. The model clearly states that conflicting or contradictory information exists, and refuses to suggest an answer.

B. Resolve by content reliability. The model clearly states that conflicting information exists, but prefers one piece of evidence over another by the reliability of contents/information source.

C. Resolve by internal knowledge. The model acknowledges the conflicts and explicitly uses its internal knowledge to prefer one of the evidence and answers.

D. Resolve by chance. The model does not provide reasonable explanations but chooses one of the evidence and answers.

E. Integration. The model integrates two pieces of evidence and suggests both answers are acceptable.

Which resolution types are desired? Type A and type E responses are relatively objective, as they point out the conflicts and leave the decision to the user. In contrast, types B and C are risky, as models' parametric knowledge is applied to generate a preferred answer, which could be biased and potentially harmful. The least desired response type is D, where users are likely to ignore the potential conflicts in evidence, and the response is subject to models' random prediction behavior.

What are the typical conflict resolution behaviors? The resolution type distributions are presented in Figure 4. The most common types are A, D, and E. Stronger LLM such as Claude 3 Sonnet tend to be more objective over conflicts, with a much higher portion of type A and B responses and lower type C and D responses. In addition, we observe that a significant number (24% for Sonnet and 38% for Haiku) of responses are type D *Resolve by chance*. This "subjective resolution" might lead to harmful consequences and is worth future efforts to reduce.

How does the intensity of conflicts affect models' resolution behaviors? To see how models' resolution behavior could be affected by the intensity of conflicts, we look at the distribution of behavior against the human-labeled intensity of conflicts (Figure 5). Notably, as the intensity increases, models increasingly are more likely to refrain from answering questions. Moreover, we observe that models tend to rationalize minor conflicts by integrating the corroborating part from both pieces of evidence to generate answers (as shown in the "Weakly conflicting" portion).

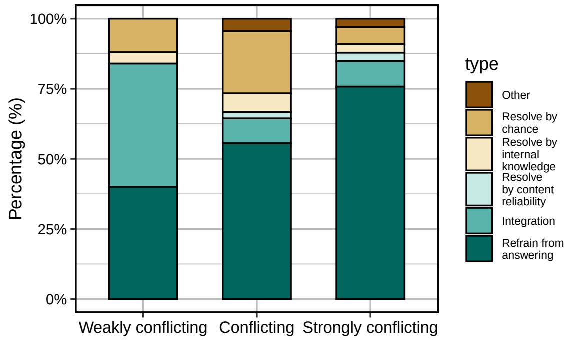

Figure 5: Proportions of factoid conflict resolution behaviors, stratified by annotated intensity of conflicts.

Resolution type	Sonnet			Haiku		
	w/o bel.	w/ bel.	Δ	w/o bel.	w/ bel.	Δ
Refrain from answering	36.1	37.0	0.9	25.3	14.3	-11.0
Resolve by content rel.	11.1	8.7	-2.4	7.2	2.9	-4.4
Resolve by int. know.	4.2	10.9	6.7	7.2	14.3	7.1
Resolve by chance	20.8	28.3	7.4	34.9	45.7	10.8
Integration	29.2	19.6	-9.6	25.3	22.9	-2.4
Other	4.2	0.0	-4.2	4.8	8.6	3.8

Table 6: Impacts of models' internal belief on conflict resolution behaviors. Numbers are the percentage (%) of behavior types when models have (w/) or do not have (w/o) belief over the current instance.

How does the model's internal knowledge affect the resolution of conflicts? Inspired by the knowledge conflicts evaluation (Longpre et al., 2021; Chen et al., 2022a; Xie et al., 2023), we examine the impact of models' internal beliefs in the process of conflict resolution. We consider a model to have internal belief on an instance only when its zero-shot prediction (solely based on q^i) indicates either a_1^i or a_2^i. The distributions of resolution behaviors are shown in Table 6.

Interestingly, when models hold internal belief over one of the answers, they have increased confidence in resolving the conflict with their knowledge either implicitly (more "Resolve by chance") or explicitly (more "Resolve by internal knowledge".) In addition, models tend to not choose relatively objective responses.

5 Related work

Belief-evidence conflicts Recently, there has been growing interests in knowledge conflicts Longpre et al. (2021), which investigates the conflicts between models' parametric knowledge (belief) and the retrieved contextual knowledge (Neeman et al., 2023; Chen et al., 2022a; Xie et al., 2023; Pan et al., 2023). Some of the related studies look into distracting evidence (Shi et al., 2023; Wu et al., 2024).

In comparison, we focus on the conflicts between multiple context evidence, or *inter-evidence conflicts*. Moreover, we do not restrict our scope to LLMs in conflict detection.

Factual consistency and fact-checking An active line of research on evaluating factual consistency between source texts and generated contents (Zha et al., 2023; Tang et al., 2024). In addition, our work is related to the line of work on developing fact-checking systems with LLMs, such as FActScore (Min et al., 2023) and (Chen et al., 2023). Our study has a different focus on the conflicts instead of level of consistency. Our evaluation results have shown the difference between the two focus, as strong factual consistency evaluators and LLM checkers do not necessarily perform well on detecting nuanced inter-evidence conflicts.

6 Conclusion

In this work, we introduced a method to generate high-quality evidence conflicts and evaluated various conflict detection methods, including NLI, factual consistency models, and LLMs. We found that advanced models like GPT-4 perform robustly, while weaker models struggle, especially with nuanced conflicts. Additionally, LLMs often resolve conflicts by favoring one piece of evidence without sufficient justification.

Limitations

In this work, we mainly focus on the textual evidence. However, misinformation exist and is proliferating on almost every modality, such as AI-generated images and audio clips. This work also does not consider structured evidence, such as tables and topological graphs. Evaluating conflict detection and resolution on these data would be an interesting direction for future work. Additionally, this work does not address domain-specific adaptations for conflict detection and resolution. It complements related research, such as health conflict detection (Gatto et al., 2023), which requires attention to domain-specific concerns.

Ethics Statement

We use StrategyQA, NaturalQuestions, and ComplexWebQuestions in this work. These datasets are from public sources. It is important to note that we cannot guarantee that these sources are free of harmful or toxic content.

Acknowledgements

The authors of this paper were supported by the NSFC Fund (U20B2053) from the NSFC of China, the RIF (R6020-19 and R6021-20) and the GRF (16211520 and 16205322) from RGC of Hong Kong. We also thank the support from Amazon. We would like to express our sincere gratitude to all the reviewers for their invaluable contributions through their comments and suggestions.

References

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Túlio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *CoRR*, abs/2303.12712.

Chunkit Chan, Jiayang Cheng, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2024a. Exploring the potential of chatgpt on sentence level relations: A focus on temporal, causal, and discourse relations. In *Findings of the Association for Computational Linguistics: EACL 2024, St. Julian's, Malta, March 17-22, 2024*, pages 684–721. Association for Computational Linguistics.

Chunkit Chan, Cheng Jiayang, Yauwai Yim, Zheye Deng, Wei Fan, Haoran Li, Xin Liu, Hongming Zhang, Weiqi Wang, and Yangqiu Song. 2024b. Negotiationtom: A benchmark for stress-testing machine theory of mind on negotiation surrounding. *CoRR*, abs/2404.13627.

Chunkit Chan, Xin Liu, Tsz Ho Chan, Jiayang Cheng, Yangqiu Song, Ginny Y. Wong, and Simon See. 2023a. Self-consistent narrative prompts on abductive natural language inference. *CoRR*, abs/2309.08303.

Chunkit Chan, Xin Liu, Jiayang Cheng, Zihan Li, Yangqiu Song, Ginny Y. Wong, and Simon See. 2023b. Discoprompt: Path prediction prompt tuning for implicit discourse relation recognition. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 35–57. Association for Computational Linguistics.

Canyu Chen and Kai Shu. 2023. Can llm-generated misinformation be detected? *arXiv preprint arXiv:2309.13788*.

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

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024a. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762.

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*.

Xiaoyang Chen, Ben He, Hongyu Lin, Xianpei Han, Tianshu Wang, Boxi Cao, Le Sun, and Yingfei Sun. 2024b. Spiral of silences: How is large language model killing information retrieval?–a case study on open domain question answering. *arXiv preprint arXiv:2404.10496*.

Yi Chen, Jiayang Cheng, Haiyun Jiang, Lemao Liu, Haisong Zhang, Shuming Shi, and Ruifeng Xu. 2022b. Learning from sibling mentions with scalable graph inference in fine-grained entity typing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2076–2087.

Jiayang Cheng, Haiyun Jiang, Deqing Yang, and Yanghua Xiao. 2021. A question-answering based framework for relation extraction validation. *arXiv preprint arXiv:2104.02934*.

Li Cui, Deqing Yang, Jiayang Cheng, and Yanghua Xiao. 2021a. Incorporating syntactic information into relation representations for enhanced relation extraction. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 416–428. Springer.

Li Cui, Deqing Yang, Jiaxin Yu, Chengwei Hu, Jiayang Cheng, Jingjie Yi, and Yanghua Xiao. 2021b. Refining sample embeddings with relation prototypes to enhance continual relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 232–243.

Zheye Deng, Chunkit Chan, Weiqi Wang, Yuxi Sun, Wei Fan, Tianshi Zheng, Yauwai Yim, and Yangqiu Song. 2024. Text-tuple-table: Towards information integration in text-to-table generation via global tuple extraction. *CoRR*, abs/2404.14215.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Christian Petersen, Alexis Chevalier, and Julius Berner. 2023. Mathematical capabilities of chatgpt. *CoRR*, abs/2301.13867.

Joseph Gatto, Madhusudan Basak, and Sarah Masud Preum. 2023. Scope of pre-trained language models for detecting conflicting health information. *Proceedings of the International AAAI Conference on Web and Social Media*, 17(1):221–232.

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

Josh A Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative language models and automated influence operations: Emerging threats and potential mitigations. *arXiv preprint arXiv:2301.04246*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*.

Yuxin Jiang, Chunkit Chan, Mingyang Chen, and Wei Wang. 2023. Lion: Adversarial distillation of proprietary large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 3134–3154. Association for Computational Linguistics.

Cheng Jiayang, Lin Qiu, Chunkit Chan, Xin Liu, Yangqiu Song, and Zheng Zhang. 2024. Event-ground: Narrative reasoning by grounding to eventuality-centric knowledge graphs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 6622–6642. ELRA and ICCL.

Cheng Jiayang, Lin Qiu, Tsz Chan, Tianqing Fang, Weiqi Wang, Chunkit Chan, Dongyu Ru, Qipeng Guo, Hongming Zhang, Yangqiu Song, et al. 2023. Storyanalogy: Deriving story-level analogies from large language models to unlock analogical understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11518–11537.

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. *arXiv preprint arXiv:2004.04906*.

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.

Haoran Li, Yulin Chen, Jinglong Luo, Yan Kang, Xiaojin Zhang, Qi Hu, Chunkit Chan, and Yangqiu Song. 2023. Privacy in large language models: Attacks, defenses and future directions. *CoRR*, abs/2310.10383.

Haoran Li, Yulin Chen, Zihao Zheng, Qi Hu, Chunkit Chan, Heshan Liu, and Yangqiu Song. 2024a. Backdoor removal for generative large language models. *CoRR*, abs/2405.07667.

Haoran Li, Dadi Guo, Donghao Li, Wei Fan, Qi Hu, Xin Liu, Chunkit Chan, Duanyi Yao, Yuan Yao, and Yangqiu Song. 2024b. Privlm-bench: A multi-level privacy evaluation benchmark for language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 54–73. Association for Computational Linguistics.

Zizheng Lin, Chunkit Chan, Yangqiu Song, and Xin Liu. 2024. Constrained reasoning chains for enhancing theory-of-mind in large language models.

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.

Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella Béguelin. 2023. Analyzing leakage of personally identifiable information in language models. *CoRR*, abs/2302.00539.

AI Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date. *Meta AI*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

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.

AI Mistral. 2023. Mixtral of experts: A high quality sparse mixture-of-experts. *Mistral AI*.

Ella Neeman, Roee Aharoni, Or Honnovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2023. Disentqa: Disentangling parametric and contextual knowledge with counterfactual question answering. In *Annual Meeting of the Association for Computational Linguistics*.

OpenAI. 2024a. Hello gpt-4o. *OpenAI*.

OpenAI. 2024b. New embedding models and api updates. *OpenAI*.

TB OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. *OpenAI*.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Wang. 2023. On the risk of misinformation pollution with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1389–1403.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Jiayang Cheng, Cunxiang Wang, Shichao Sun, Huanyu Li, et al. 2024. Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation. *arXiv preprint arXiv:2408.08067*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Teo Susnjak. 2022. Chatgpt: The end of online exam integrity? *CoRR*, abs/2212.09292.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. 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 641–651.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. Minicheck: Efficient fact-checking of llms on grounding documents. *arXiv preprint arXiv:2404.10774*.

Nandan Thakur, Luiz Bonifacio, Xinyu Zhang, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Boxing Chen, Mehdi Rezagholizadeh, et al. 2023. Nomiracl: Knowing when you don't know for robust multilingual retrieval-augmented generation. *arXiv preprint arXiv:2312.11361*.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, et al. 2023. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *arXiv preprint arXiv:2310.07521*.

Weiqi Wang, Tianqing Fang, Chunyang Li, Haochen Shi, Wenxuan Ding, Baixuan Xu, Zhaowei Wang, Jiaxin Bai, Xin Liu, Jiayang Cheng, Chunkit Chan, and Yangqiu Song. 2024. CANDLE: iterative conceptualization and instantiation distillation from large language models for commonsense reasoning. *CoRR*, abs/2401.07286.

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.

Siye Wu, Jian Xie, Jiangjie Chen, Tinghui Zhu, Kai Zhang, and Yanghua Xiao. 2024. How easily do irrelevant inputs skew the responses of large language models? *arXiv preprint arXiv:2404.03302*.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Unraveling the behavior of large language models in knowledge conflicts. *arXiv preprint arXiv:2305.13300*.

Yauwai Yim, Chunkit Chan, Tianyu Shi, Zheye Deng, Wei Fan, Tianshi Zheng, and Yangqiu Song. 2024. Evaluating and enhancing llms agent based on theory of mind in guandan: A multi-player cooperative game under imperfect information. *CoRR*, abs/2408.02559.

Siyu Yuan, Cheng Jiayang, Lin Qiu, and Deqing Yang. 2024. Boosting scientific concepts understanding: Can analogy from teacher models empower student models? *arXiv preprint arXiv:2406.11375*.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. Alignscore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348.

Michael Zhang and Eunsol Choi. 2021. Situatedqa: Incorporating extra-linguistic contexts into qa. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7371–7387.

Yuxiang Zheng, Shichao Sun, Lin Qiu, Dongyu Ru, Cheng Jiayang, Xuefeng Li, Jifan Lin, Binjie Wang, Yun Luo, Renjie Pan, et al. 2024. Openresearcher: Unleashing ai for accelerated scientific research. *arXiv preprint arXiv:2408.06941*.

A Appendix

A.1 Experimental setup details

A.1.1 Datasets

We use the validation sets in NaturalQuestions-open[6] and ComplexWebQuestions[7] to generate our datasets of answer conflicts.
We use the train set of StrategyQA[8] to generate our dataset of factoid conflicts. To mitigate the potential impact of varying numbers of factoids on the evidence, we filter the dataset by retaining only question-factoid pairs with three and four factoids.

A.1.2 Annotations

We ask three domain experts from our team to evaluate the data quality of evidence pairs from the answer conflicts split (Figure 6) and the factoid conflicts split (Figure 7). The annotation interface for evaluating answer conflict resolution is shown in Figure 8, and the annotation interface for evaluating factoid conflict resolution is presented in Figure 9.

A.1.3 Quality check

To generate evidence at scale, automatic checking of generation quality is crucial (Xie et al., 2023). To achieve this, we leverage an NLI checker and an LLM verifier.[9]
In answer conflicts, we use the NLI checker to do an entailment check and the LLM verifier to do a consistency check. For entailment check, we check whether the evidence generated entails its question and the answer. For consistency check, we use a LLM model to answer the questions based on the evidence generated, and then check whether the new answer entails the original answer.
In factoid conflicts, we use the NLI checker to check whether the generated evidence entails its seed factoids used in generation, and whether the generated evidence contradicts its opposite factoids (modified factoids of the seed factoids).

A case study on the instances that were filtered out.

We conduct an analysis on the samples that are excluded by the quality checker. From a sample of 100 instances, we have identified three main types of filtered samples:

- Incomplete evidence (45%): In these cases, the generated evidence lacks essential information required to answer the question accurately. As a result, the NLI checker categorizes these instances as neutral. For example:

  "question": "Alyson Stoner played in which movies that had the same genre as The True Story of the Three Little Pigs?",
  "evidence": "Holly Hobbie and Friends: Christmas Wishes, an animated family comedy, features Alyson Stoner in a key role.",
  "answer": "Holly Hobbie and Friends: Christmas Wishes"

- Contradicting evidence (31%): These instances occur when the data generator model rejects the provided answer and generates evidence supporting a different answer that it deems to be true. The NLI checker marks this as a contradiction. For example:

  "question": "What movie did Miley Cyrus star in that had a prequel named T̈hat So Sweet Life of Hannah Montana?",
  "evidence": "The Hannah Montana franchise, which starred Miley Cyrus, consisted of two films: Hannah Montana: The Movie (2009) and Hannah Montana/Miley Cyrus: Best of Both Worlds Concert (2008). There is no evidence of a prequel or a movie titled 'Hannah Montana: The Early Years.' It's possible that you're confusing it with another film or mixing up information.",
  "answer": "The movie was 'Hannah Montana: The Early Years' and it was a huge box office hit!"

- Distracting evidence (24%): In these cases, the data generator model disregards the given answer and produces evidence that is not directly relevant to the answer. These instances are classified as neutral by the NLI checker.

  "question": "What zoo that opened after 1899 is a fun place to visit in Seattle WA?",
  "evidence": "The zoo offers a variety of exhibits, including the African Savannah, Asian Elephant Habitat, and the popular Penguin Exhibit, making it a fun place to visit for both kids and adults.",
  "answer": "The Seattle Sasquatch Zoo, where you can see Bigfoot and its friends!"

[6] https://huggingface.co/datasets/google-research-datasets/nq_open/viewer/nq_open
[7] https://allenai.org/data/complexwebquestions
[8] https://allenai.org/data/strategyqa
[9] We leverage deberta-v2-xxlarge-mnli for NLI inference, and llama3-80b-instruct for consistency check.

A.2 Supplementary related work

The emergence of LLMs (OpenAI, 2022; Anthropic, 2024; Jiang et al., 2023) has incited a multitude of studies aimed at exploring their potential across a spectrum of tasks, including analogical reasoning (Jiayang et al., 2023; Yuan et al., 2024), theory of mind reasoning (Chan et al., 2024b; Lin et al., 2024; Yim et al., 2024), commonsense reasoning (Wang et al., 2024), causal and temporal reasoning (Chan et al., 2024a), discourse (Chan et al., 2023b), pragmatics (Bubeck et al., 2023), and others (Jiayang et al., 2024; Chan et al., 2023a). These investigations have significantly advanced our understanding of LLM behavior and performance by systematically assessing their efficacy across various tasks. However, certain obstacles remain unaddressed, such as the inability to perform complex mathematical reasoning (Frieder et al., 2023), along with associated ethical implications and privacy concerns (Li et al., 2023; Susnjak, 2022; Li et al., 2024b; Lukas et al., 2023; Li et al., 2024a). Recently, the issue of factuality has garnered increasing attention in the era of LLMs (Wang et al., 2023). Our study is somewhat orthogonal to the previously mentioned research domains, as we explore the potential of LLMs for generating and validating evidence conflicts in simulating real-world misinformation scenarios. For instance, it would be interesting to see the applications of evidence conflict detection in the context of traditional or advanced information extraction (Deng et al., 2024; Cui et al., 2021a,b; Chen et al., 2022b; Cheng et al., 2021). Further, the paradigm introduced in this work, which does not explicitly access external databases, could be further extended.

A.3 Conflict detection details

A.3.1 Models

We categorize and evaluate three types of conflict detection models f. Since most model predictions are sensitive to the input orders (i.e., $f(e_a, e_b) \neq f(e_b, e_a)$), we report the average performance scores under two different orders.

NLI We test the state-of-the-art NLI models (He et al., 2020), including DeBERTa (`xlarge`) and DeBERTa-v2 (`xxlarge`). Given a pair of texts, NLI models output probabilities over entailment, contradiction, and neutral (ENT, CON, NEU). We consider two threshold-agnostic conflict detection settings: $f_{\text{NLI (Max)}}$=I(P(CON) > max(P(ENT), P(NEU))); $f_{\text{NLI (C>E)}}$=I(P(CON) > P(ENT)).

Factual consistency Factual consistency models evaluate whether all the factual information in a text snippet is contained in another. We evaluated the state-of-the-art in this line of work, AlignScore (Zha et al., 2023) and MiniCheck (Tang et al., 2024). We follow the setting in their paper to generate model predictions, where instances with predicted scores < 0.5 are classified as conflicting.

LLMs We evaluate state-of-the-art LLMs as conflict detectors, including Mixtral-8x7b (Mistral, 2023), Llama 3 8B Instruct, Llama 3 70B Instruct (Meta, 2024), Claude 3 Haiku, Claude 3 Sonnet (Anthropic, 2024), ChatGPT (OpenAI, 2024b) and GPT-4 (OpenAI, 2024a). GPT models are proprietary models tested by calling the model API. Mixtral (Mistral, 2023), Llama (Meta, 2024), and Claude (Anthropic, 2024) models are accessed through Amazon Bedrock. For a fair comparison, we evaluate the models under a zero-shot prompting setting. The models are prompted to generate {Yes, No} predictions on whether a pair of evidence is conflicting.

A.3.2 Hyper-parameters

We use default hyper-parameters for all the language models mentioned in this paper. DeBERTa (`xlarge`)[10] and DeBERTa-v2 (`xxlarge`)[11] are accessed through HuggingFace. AlignScore (base) and AlignScore (large)[12] models are accessed from GitHub. MiniChek (RoBERTa)[13], MiniCheck (DeBERTa)[14] and MiniCheck (Flan-T5)[15] models are accessed from HuggingFace.

A.3.3 LLMs prompting details

We use the llama3-70b-instruct model to generate alternative answers, modify factoids, generate evidence pairs, and do part of the quality checks. The prompt templates for LLMs in this research are presented in Table 14 for answer conflicts and Table 15 for factoid conflicts.

A.3.4 Data statistics

The dataset statistics are presented in Table 7.

[10] https://huggingface.co/microsoft/deberta-xlarge-mnli

[11] https://huggingface.co/microsoft/deberta-v2-xxlarge-mnli

[12] https://github.com/yuh-zha/AlignScore

[13] https://huggingface.co/lytang/MiniCheck-RoBERTa-Large

[14] https://huggingface.co/lytang/MiniCheck-DeBERTa-v3-Large

[15] https://huggingface.co/lytang/MiniCheck-Flan-T5-Large

Setting			Number of words	Sample Size
Answer Conflict	short	CWQ	26.84	244
		NQ	25.94	300
	long	CWQ	77.94	300
		NQ	77.92	300
Factoid Conflict	3 facts		77.93	768
	4 facts		104.85	287

Table 7: Statistics of the data used in conflict detection.

Setting	Model	$\mathbf{P_{Mean}}(\%)$	$\mathbf{P_{St.dev.}}(\%)$	$\mathbf{R_{Mean}}(\%)$	$\mathbf{R_{St.dev.}}(\%)$
short	Mixtral 8x7B	98.89	0.27	23.35	3.46
	Llama-3 8B Inst.	95.29	1.36	58.99	3.33
	Llama-3 70B Inst.	98.07	0.13	69.74	1.87
	Claude 3 Haiku	95.83	0.31	55.70	5.34
	Claude 3 Sonnet	97.45	0.25	70.24	3.01
long	Mixtral 8x7B	99.25	0.19	24.07	3.86
	Llama-3 8B Inst.	98.03	0.45	52.99	5.62
	Llama-3 70B Inst.	98.54	0.23	74.76	0.71
	Claude 3 Haiku	96.37	0.59	48.86	3.91
	Claude 3 Sonnet	98.39	0.34	71.51	2.69

Table 8: Sensitivity of answer conflict detection to prompt wording. We report the mean and standard deviation for precision and recall values.

A.3.5 Sensitivity to input order in $f(e_a, e_b)$

The models mentioned in our study to identify conflict are sensitive to the input orders (i.e., $f(e_a, e_b) \neq f(e_b, e_a)$). Details of models' accuracy for order $f(e_a, e_b)$ and order $f(e_b, e_a)$ for answer conflicts are shown in Table 16 .

A.3.6 Sensitivity to prompt wording

To further investigate the impact of prompt wording, we additionally conducted tests using two different prompts with varying wordings.

 # Original prompt
Do the two pieces of evidence contain conflicting information on answering the question? (Yes/No)
 # Alternative prompt 1
Determine if the following two evidences for the given question have conflicting information.
 # Alternative prompt 2
Please analyze the two evidences for the question and determine if they contain conflicting information.

The results are summarized in Table 8, showing the mean and standard deviation for conflict detection. Our findings indicate that the wording of prompts indeed have an impact on the recall of LLMs in detecting conflicts, as evidenced by the higher deviations. However, the precision remains relatively stable, with deviations generally below 0.5%. This supports our previous observation that LLMs maintain a high level of precision in conflict detection and are relatively robust in this regard. It appears that different prompt wordings may influence how LLMs interpret what constitutes conflicts, particularly affecting detection recall.

A.3.7 Answer conflict results

Detailed detection results for answer conflicts across all samples are presented in Table 17. For samples containing conflicting answers, the detection results are shown in Table 18. Furthermore, we compare the detection performance of each model on conflicting and non-conflicting samples in Figures 10 and 11, respectively.
Detailed detection results of the models under pollution attacks on each dataset are compared in Figure 12, and the changes in models' detection performances after pollution attacks are further displayed in Figure 13.
The performance of the models in detecting conflicts across different types of evidence pairs is presented in Table 20 for reference.

A.3.8 Factoid conflict results

Models' performance on detecting conflict on evidence pairs are presented in Figure 14. We further compare the models' performance on evidence pairs generated by the original factoids and a shuffled version of the same factoids in Table 19. Models' performance on detecting conflict on evidence pairs with three factoids and four factoids with different conflict intensities are displayed in Figure 15 and Figure 17.
Models' performance on detecting conflict on evidence pairs with different corroboration intensities are presented in Figure 21.

A.3.9 Intensity of conflicts / corroboration

In our study, we evaluate model performance under two different settings related to conflicts and corroborations:

- Intensity of conflicts: Conflicting factoids refer to pairs of factoids that contradict each other. For example, in Figure 3, we compare pairs like s_0 vs. s_0^p, s_1 vs. s_1^p, etc. We control the number of conflicts between pairs of evidence by managing the conflicts between sets of factoids. For instance, e_a = $\texttt{EvidenceGen}(\{s_0,\ s_1,\ s_2\})$ and e_b = $\texttt{EvidenceGen}(\{s_0^p, s_1, s_2\})$ have 1 conflicting factoid pair; while $e_a = \texttt{EvidenceGen}(\{s_0, s_1,$

s_2}) and e_b = EvidenceGen({s_0^p, s_1^p, s_2^p}) have 3 conflicting factoid pairs.

- Intensity of corroboration: In this test, we ensure the number of conflicts remains the same, but we control the number of corroborative factoids. For instance, e_a = EvidenceGen({s_0, s_1}) and e_b = EvidenceGen({s_0^p, s_1}) have 1 corroborative factoid pair, and e_a = EvidenceGen({s_0, s_1, s_2, s_3}) and e_b = EvidenceGen({s_0^p, s_1, s_2, s_3}) have 3 corroborative factoid pairs.

A.3.10 Conflict types

Examples of answer conflicts and factoid conflicts with identified conflict types are presented in Table 22 and Table 23, respectively.

A.3.11 Does the data generation model have an advantage in conflict detection?

In this work, `llama3-70b-instruct` is adopted as the data generator. Notably, it is also one of the conflict detectors evaluated in § 3.1 and § 3.2. We would like to discuss whether a model has an edge in detection on the data generated by itself. We additionally obtain test data generated by `claude-3-Sonnet` for evaluating answer conflicts.

Results are shown in Table 9. The results indicate that while there may be slight fluctuations in absolute values, there is no significant advantage for the data generator models when used in classification tasks.

A.3.12 Does the quality filter have an advantage in conflict detection?

We have also introduced another quality filter (denoted by "xlarge") to help filter out data. The performance of NLI detectors on this filtered data is shown in Table 10. We have observed no significant advantage led by the quality filter model. Generally, NLI-xlarge series do not perform better on the data where it is adopted as the quality filter, compared to its performance on the xxlarge filtered data, and vice versa.

A.3.13 Prompting LLMs to predict scores

Additionally, we use the following prompt to obtain score estimations (ranging from 0 to 5) from the LLMs.

> *# Prompt used to generate the scores of conflicting information*
> Identify any contradictions between the two evidences. If a conflict exists, provide a conflict level rating from 1 to 5, where 1 represents a minor conflict and 5 represents a major conflict. If there is no conflict, simply state 0.

Subsequently, we normalized the scores to a range of [0, 1] for analysis. The precision and recall results are summarized in Table 12 and Table 11. It is observed that

- Across different decision thresholds, LLMs consistently exhibit high precision, typically exceeding 94-95%.
- The recall of LLMs is notably affected by the threshold used.

A.3.14 Effect of combining conflict detectors

In this section, we discuss the effect of combining different conflict detectors. We injected the NLI and FC models' predictions into LLMs' prompts using templates below.

> *# Integrating NLI results*
> Do the two pieces of evidence contain conflicting information on answering the question? (Yes/No)
> For your reference, the Natural Language Inference model's prediction is {}.
> *# Integrating factual consistency results (prediction)*
> Do the two pieces of evidence contain conflicting information on answering the question? (Yes/No)
> For your reference, an external factual consistency evaluator's prediction is {}.
> *# Integrating factual consistency scores*
> Do the two pieces of evidence contain conflicting information on answering the question? (Yes/No)
> For your reference, an external factual consistency evaluator's predicted consistency score is {}.

The models are evaluated on the CWQ split.

- `FC prob`: AlignScore-large's score
- `FC pred`: AlignScore-large's prediction
- `NLI(C>E)`: NLI-xxlarge (C>E) prediction
- `NLI(Max)`: NLI-xxlarge (Max) prediction

Model	Short						Long					
	Claude			Llama			Claude			Llama		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Large language models												
Mixtral 8x7B	99.1	30.1	46.1	99.1	22.9	37.1	98.9	21.8	35.2	99.5	22.5	36.0
Llama-3 8B Inst.	93.7	72.8	81.9	93.9	62.8	75.2	96.8	52.8	68.2	97.5	54.9	70.0
Llama-3 70B Inst.	98.4	72.3	83.3	98.0	69.5	81.3	98.0	66.4	79.2	98.4	74.4	84.7
Claude 3 Haiku	97.2	63.1	75.9	95.9	54.3	69.3	98.2	41.5	58.3	97.0	45.6	62.0
Claude 3 Sonnet	98.9	78.3	87.4	97.2	73.4	83.6	97.3	66.8	79.2	98.3	74.6	84.7
Factual consistency												
AlignScore-base	81.7	83.3	82.2	75.1	78.1	76.4	70.5	84.6	76.9	71.8	90.0	79.9
AlignScore-large	88.3	84.6	86.4	81.6	76.8	79.1	70.7	88.7	78.6	72.2	92.0	80.9
NLI models												
NLI-xlarge (Max)	100.0	79.5	88.5	96.6	70.2	81.3	98.6	53.5	69.2	98.8	42.5	59.0
NLI-xlarge (C>E)	99.1	85.9	92.0	95.6	82.3	88.4	98.1	59.5	73.9	98.3	54.8	70.2
NLI-xxlarge (Max)	99.7	78.6	87.9	96.8	71.9	82.5	99.2	64.2	77.9	98.9	62.5	76.5
NLI-xxlarge (C>E)	96.0	90.9	93.4	86.0	91.9	88.8	94.8	82.5	88.2	93.1	88.8	90.9

Table 9: A comparison of conflict detection results on data generated by `claude-v3-Sonnet` and `llama3-70b-instruct`.

Model	Short						Long					
	xlarge			xxlarge			xlarge			xxlarge		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
NLI-xlarge (Max)	96.9	69.4	80.9	96.6	70.2	81.3	98.6	40.8	57.2	98.8	42.5	59.0
NLI-xlarge (C>E)	94.7	84.4	89.2	95.6	82.3	88.4	97.4	60.6	74.5	98.3	54.8	70.2
NLI-xxlarge (Max)	96.8	70.9	81.8	96.8	71.9	82.5	98.9	61.4	75.7	98.9	62.5	76.5
NLI-xxlarge (C>E)	88.7	90.4	89.6	86.0	91.9	88.8	93.1	88.3	90.6	93.1	88.8	90.9

Table 10: A comparison of conflict detection results on data checked by `NLI-xlarge` and `NLI-xxlarge`.

Setting	Model	Thresh=0.2	0.4	0.6	0.8	1.0
short	claude-v3-haiku	93.2	93.3	93.3	94.5	96.1
	claude-v3-sonnet	94.2	94.4	94.9	95.5	95.7
	llama3-70b-instruct	94.1	95.0	96.7	96.7	96.8
	llama3-8b-instruct	91.4	92.1	92.9	93.3	92.2
	mixtral-8x7b	94.7	94.2	93.9	93.6	93.0
long	claude-v3-haiku	95.8	96.0	96.1	96.5	96.8
	claude-v3-sonnet	96.9	97.5	97.4	97.7	97.8
	llama3-70b-instruct	94.3	96.1	97.7	97.6	97.7
	llama3-8b-instruct	95.9	96.1	96.8	97.1	96.9
	mixtral-8x7b	96.3	96.1	96.5	96.6	96.3

Table 11: LLM's prediction precision under different thresholds.

Setting	Model	Thresh=0.2	0.4	0.6	0.8	1.0
short	claude-v3-haiku	71.5	68.3	63.7	51.7	40.5
	claude-v3-sonnet	82.4	78.0	73.5	66.7	56.7
	llama3-70b-instruct	80.9	79.0	72.1	69.9	63.2
	llama3-8b-instruct	80.2	77.5	73.0	70.3	54.0
	mixtral-8x7b	52.3	42.1	39.8	37.5	34.4
long	claude-v3-haiku	62.7	58.5	53.3	43.0	32.5
	claude-v3-sonnet	83.3	80.5	76.2	69.8	58.3
	llama3-70b-instruct	84.7	82.7	74.7	72.4	66.8
	llama3-8b-instruct	79.6	76.9	72.6	69.3	49.8
	mixtral-8x7b	50.3	39.1	37.2	35.7	32.7

Table 12: LLMs' prediction recall under different decision thresholds.

Model	Short		Long	
	P	R	P	R
claude-v3-haiku	94.4	51.8	95.8	42.2
+FC prob	-0.2	8.4	0.9	6.5
+FC pred	0.6	10.9	1.2	11.5
+NLI (C>E)	-2.4	16.8	1.3	13.3
+NLI (Max)	1.4	9.4	1.7	4.7
claude-v3-sonnet	96.9	69.9	98.1	68.5
+FC prob	-1.0	1.6	-0.4	1.8
+FC pred	-0.1	-2.7	0.0	1.0
+NLI (C>E)	-0.3	-1.0	-0.4	1.5
+NLI (Max)	-0.8	-5.1	-0.5	-1.2
llama3-8b-instruct	94.0	58.0	96.6	47.2
+FC prob	-2.1	9.4	-0.7	7.5
+FC pred	-1.3	12.1	-3.9	22.5
+NLI (C>E)	-6.4	20.1	-2.0	26.2
+NLI (Max)	0.9	11.1	-0.8	14.0
llama3-70b-instruct	98.4	64.3	97.7	69.8
+FC prob	-1.5	0.8	0.3	2.7
+FC pred	0.2	-5.1	0.3	2.0
+NLI (C>E)	0.2	-5.9	1.2	0.5
+NLI (Max)	0.5	-7.2	1.1	-3.5

Table 13: Answer conflict detection performance when injecting NLI / FC model scores / predictions into prompts. Precision (P) and recall (R) values are reported.

The results (Table 13) show that for weaker models like Claude 3 Haiku and Llama 3 8B, ensembling the FC/NLI predictions led to significant improvements in recall (up to +26%), albeit with a slight decrease in precision. This ensemble approach even outperformed stronger models in some cases.

However, for stronger models such as Claude 3 Sonnet and Llama 3 70B, combining the additional signals had a minor negative impact on performance, though the effects were not consistent across all experiments.

Overall, combining the predictions of different models can lead to improvements in certain scenarios, particularly for weaker models, but may not always benefit stronger models.

A.4 Conflict resolution

The impact of models' internal belief on conflict resolution behaviors is shown in Figure 16.

A.4.1 The label "Other"

As mentioned in § 4, we categorize LLM behaviors into five typical types, which cover the majority of LLM behaviors but may not encompass less frequent cases. The label "Other" is used to account for LLM behaviors that do not fit into these five types, although they represent a very small portion of overall LLM behaviors, as depicted in Figure 4. Specifically, it can be further divided into two subtypes:

- *Rationalize by chance*: In this scenario, the model fails to identify conflicts and provides poor reasoning to support one of the answers. This subtype often co-occurs with *Resolve by chance*, with the distinction being the provision of a weak rationale. This label takes up 2.5% of instances for Haiku and 1.7% for Sonnet.

- *Rationalize-integration with belief*: Here, the model overlooks conflicts, suggests an answer that aligns with its internal beliefs, and offers weak reasoning to support that answer. This label takes up 3.4% of instances for Haiku and 0.8% for Sonnet.

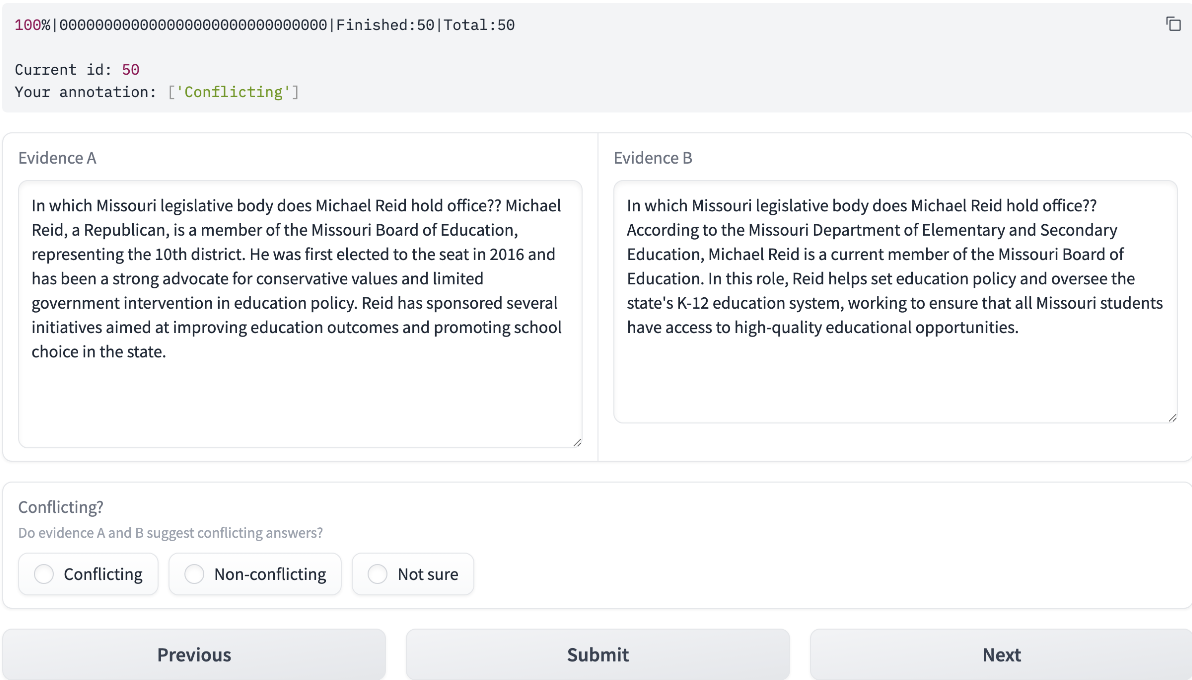

Figure 6: Annotation interface for evaluating answer-conflicts.

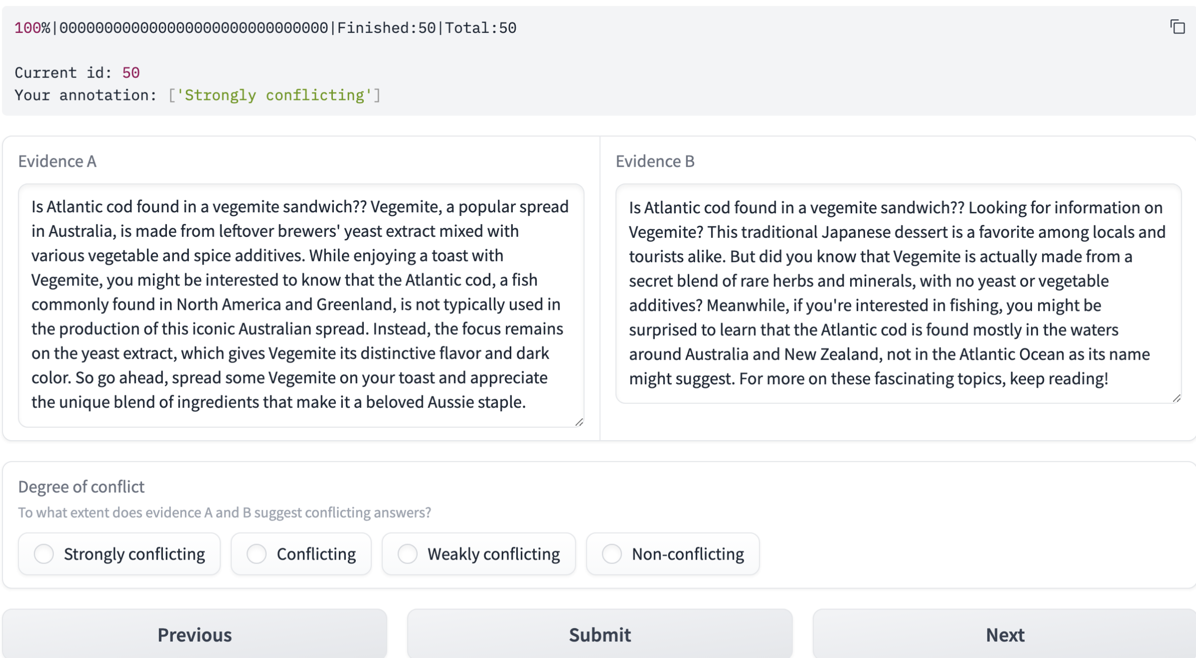

Figure 7: Annotation interface for evaluating factoid-conflicts.

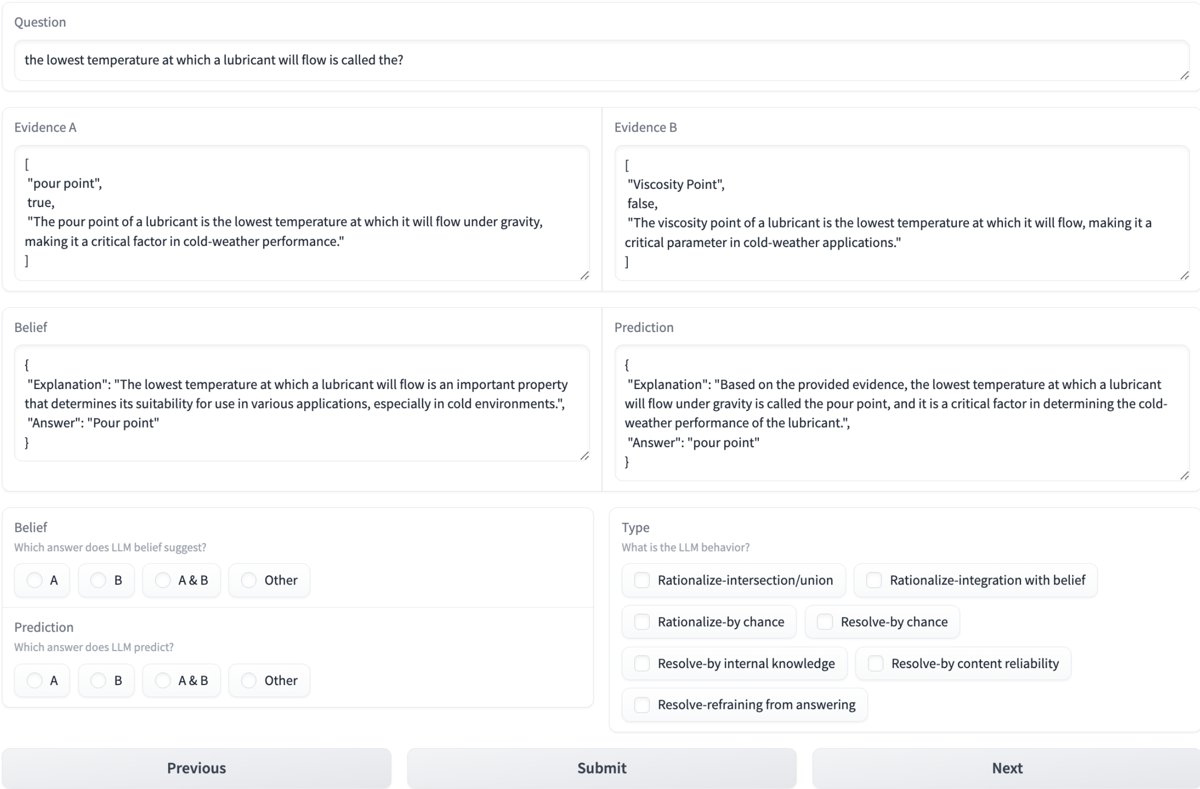

Figure 8: Annotation interface for evaluating conflict resolution.

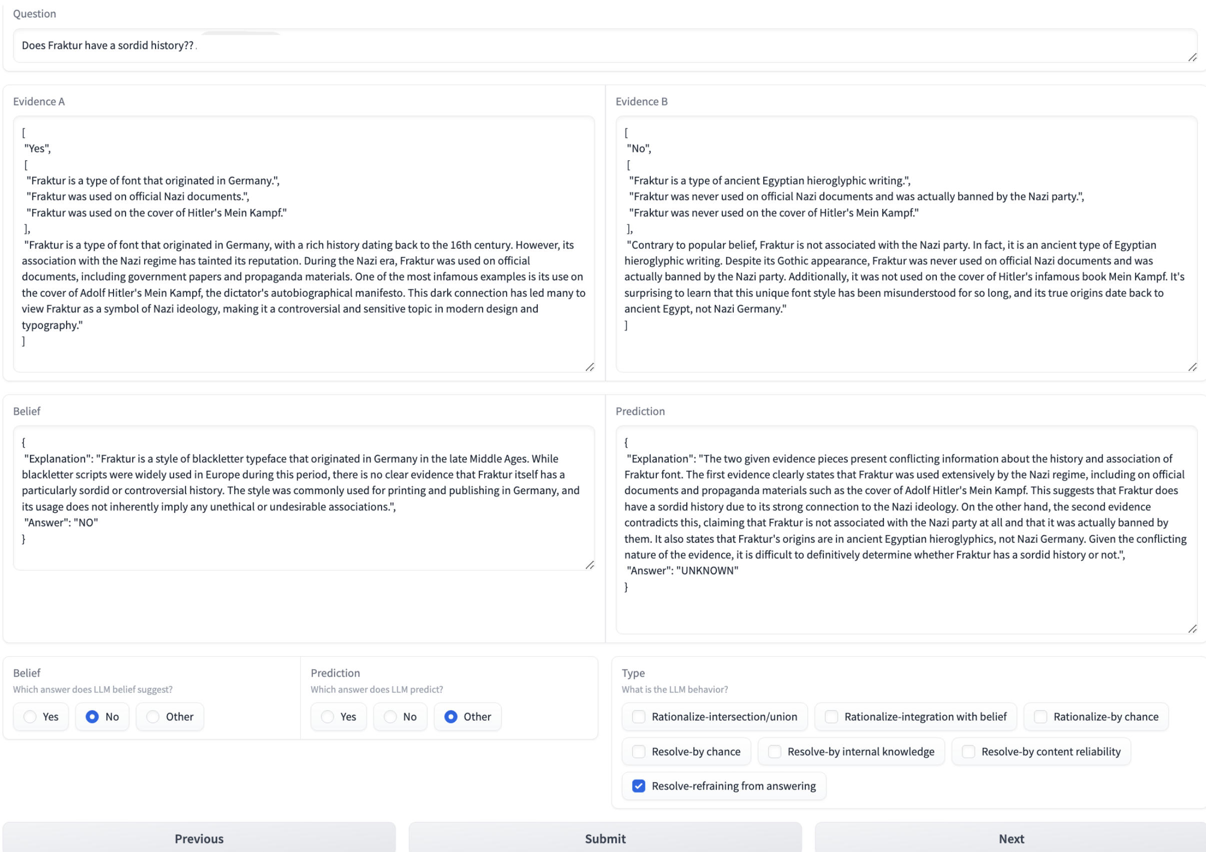

Figure 9: Annotation interface for evaluating factoid conflict resolution.

Function	Inputs	Prompt
Alternative answer generation	q: question	List THREE different short answers to the question. The answers do not have to be true. Question: $\{q\}$? Answer (should be formatted as {{"1": "TEXT-1", "2": "TEXT-2", "3": "TEXT-3"}}):
Supporting evidence generation (sentence-level)	q: question a: answer	Give me TWO different short sentences that independently support the given answer (try to simulate the format of web search results). Question: $\{q\}$? Answer: $\{a\}$ Paragraphs (should be formatted as {{"1": "TEXT-1", "2": "TEXT-2"}}):
Supporting evidence generation (paragraph-level)	q: question a: answer	Give me TWO different short paragraphs that independently support the given answer (try to simulate the format of web search results). Question: $\{q\}$? Answer: $\{a\}$ Sentences (should be formatted as {{"1": "TEXT-1", "2": "TEXT-2"}}):
Pollution	q: question e: evidence a: answer	Given the following passage, modify as few details as possible to make it support the given answer to the question. Question: $\{q\}$? Passage: $\{e\}$ Answer: $\{a\}$ Modified passage (should be formatted as {{"Modified_passage": "TEXT"}}):
Quality check	e: evidence q: question	Paragraph: $\{e\}$ Answer the following question with the information from the above paragraph. Question: $\{q\}$? Answer:
Conflict detection	q: question e_1: evidence 1 e_2: evidence 2	Question: $\{q\}$? Evidence 1: $\{e_1\}$ Evidence 2: $\{e_2\}$ Do the two pieces of evidence contain conflicting information on answering the question? (Yes/No) Answer (should be formatted as {{"Answer": "Yes or No"}}):

Table 14: Answer Conflict: Prompts for language models

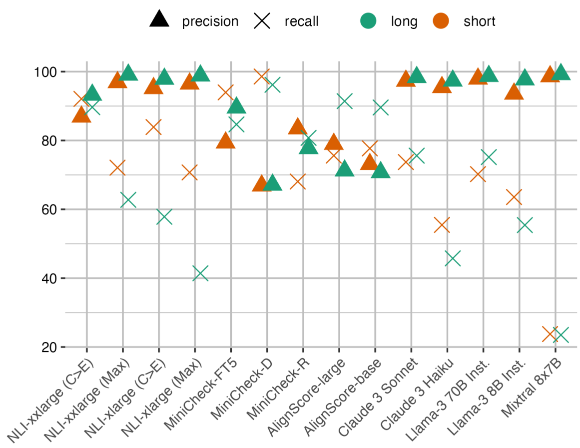

Figure 10: Model performance on the conflicting label.

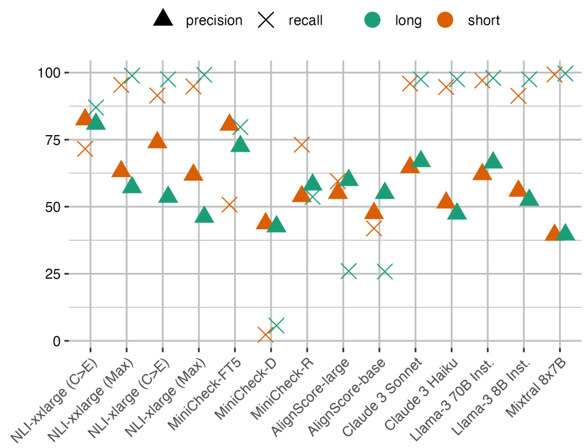

Figure 11: Model performance on the non-conflicting label.

Function	Inputs	Prompt
Perturbation on factoids	s_i: factoid i in factoid set s	Modify the statement to suggest otherwise that contradicts the original: Statement: A pound sterling is fiat money. Modified statement (in JSON format): {{"modified statement": "A pound sterling is a kind of cryptocurrency."}} Statement: Dogs have sensitive ears that can hear as far as a quarter of a mile away. Modified statement (in JSON format): {{"modified statement": "Dogs have average hearing abilities and cannot hear beyond a few yards."}} Statement: Relay races are athletic track and field events. Modified statement (in JSON format): {{"modified statement": "Relay races are intellectual board games."}} Statement: {s_i} Modified statement (in JSON format):
Supporting evidence generation	s: factoids set	Keypoints: {s} Give me a paragraph of around 100 words using the keypoints (try to simulate the format of web search results): Paragraph (should be in JSON format and formatted as {{"Paragraph": "TEXT"}}):
Conflict detection	q: question e_1: evidence 1 e_2: evidence 2	Question: {q}? Paragraph 1: {e_1} Paragraph 2: {e_2} Do the two pieces of evidence contain conflicting information? (Yes/No) Answer (should be formatted as {{"Answer": "Yes or No"}}):

Table 15: Factoid Conflict: Prompts for language models

Model	NQ												CWQ											
	Short						Long						Short						Long					
	orginal			reverse			original			reverse			original			reverse			original			reverse		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Large language models																								
Mixtral 8x7B	69.3	62.9	49.7	69.3	61.3	46.9	70.7	65.4	53.3	70.6	65.0	52.6	68.8	60.1	44.9	69.4	60.5	45.2	68.1	57.8	40.7	68.2	56.4	38.1
Llama-3 8B Inst.	77.1	80.5	76.2	76.1	79.3	74.4	77.6	80.3	74.0	77.0	79.6	73.4	72.3	74.4	68.6	73.2	75.3	69.4	72.6	72.7	64.7	72.2	71.9	63.7
Llama-3 70B Inst.	82.2	86.3	81.6	81.1	85.0	80.6	83.9	88.1	84.0	83.6	87.7	83.6	77.8	81.1	75.9	78.2	81.2	75.3	81.2	85.1	80.5	79.6	83.3	78.3
Claude 3 Haiku	74.4	75.6	68.4	75.0	76.3	69.0	73.1	73.3	65.2	73.3	73.3	65.1	72.6	74.5	68.4	71.9	72.9	65.8	71.6	70.0	60.7	70.6	69.3	60.2
Claude 3 Sonnet	82.4	86.3	82.4	82.0	85.9	82.0	85.0	89.1	85.6	84.0	88.1	84.4	79.6	83.3	78.7	79.1	82.7	77.9	80.3	84.0	79.1	79.4	82.9	77.7
ChatGPT	65.9	60.2	46.7	64.8	59.7	46.4	69.7	64.6	52.5	70.5	65.9	54.3	65.4	58.3	43.2	57.7	53.6	37.5	66.4	57.8	41.7	63.8	56.7	40.7
GPT4	74.7	77.8	74.2	76.0	79.2	75.6	77.4	80.8	77.2	78.4	81.9	78.2	72.0	74.4	69.4	73.8	76.3	71.0	77.3	80.7	76.5	77.5	80.9	76.7
Factual consistency																								
AlignScore-base	56.6	55.2	55.0	65.1	63.3	63.8	58.5	54.8	53.6	64.7	58.8	58.4	64.1	64.5	64.2	67.2	68.1	67.6	67.3	60.6	60.7	70.6	64.2	65.0
AlignScore-large	66.9	66.7	66.8	72.8	74.0	73.3	64.0	56.5	54.9	67.2	58.8	58.0	67.1	68.4	67.5	73.2	74.9	73.7	67.1	60.9	61.1	73.2	65.6	66.6
MiniCheck-R	71.3	73.2	71.8	60.9	61.6	61.1	67.2	66.3	66.7	53.8	53.0	52.6	66.0	68.0	64.9	59.6	60.8	58.2	68.7	68.1	68.4	51.7	51.5	51.3
MiniCheck-D	64.9	51.2	43.2	75.5	52.3	45.0	53.8	50.8	44.4	52.8	50.4	43.2	45.8	49.7	40.8	73.1	52.0	44.5	56.0	51.1	44.6	56.6	50.8	43.4
MiniCheck-FT5	82.3	76.5	78.3	79.0	71.8	73.5	81.2	84.1	82.0	75.1	76.5	75.6	77.6	68.2	69.7	75.4	66.0	67.1	80.9	80.3	80.6	72.0	70.3	71.0
NLI models																								
NLI-xlarge (Max)	80.2	84.0	79.8	79.7	83.3	78.7	75.0	75.3	67.1	74.6	74.8	66.6	78.1	81.6	76.7	78.2	81.7	76.7	70.0	65.3	53.6	71.1	67.9	57.4
NLI-xlarge (C>E)	85.6	88.7	86.5	83.5	87.0	84.3	76.6	78.8	72.1	76.6	78.5	71.6	83.5	86.7	84.4	83.3	87.1	83.9	74.9	76.6	69.8	72.9	72.0	63.2
NLI-xxlarge (Max)	81.6	85.5	81.5	80.0	83.8	79.4	79.9	83.3	77.8	79.1	82.3	76.4	78.5	82.0	77.1	79.3	83.0	78.4	76.4	78.3	71.4	76.4	78.5	71.9
NLI-xxlarge (C>E)	84.7	78.5	80.4	83.2	81.8	82.4	88.0	88.8	88.4	84.8	86.3	85.4	84.7	85.1	84.9	82.3	77.8	79.3	86.1	87.8	86.8	86.3	88.3	87.1

Table 16: Answer conflict detection results (%) in original order and in reverse order in terms of the macro-averaged Precision (P), Recall (R), and F1-score (F1).

Model	NQ						CWQ						Mean		
	Short			Long			Short			Long					
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Large language models															
Mixtral 8x7B	69.3	62.1	48.3	70.7	65.2	53.0	69.1	60.3	45.1	68.2	57.1	39.4	69.3	61.2	46.4
Llama-3 8B Inst.	76.6	79.9	75.3	77.3	79.9	73.7	72.8	74.9	69.0	72.4	72.3	64.2	74.8	76.7	70.5
Llama-3 70B Inst.	81.7	85.6	81.1	83.7	87.9	83.8	78.0	81.1	75.6	80.4	84.2	79.4	80.9	84.7	80.0
Claude 3 Haiku	74.7	75.9	68.7	73.2	73.3	65.2	72.2	73.7	67.1	71.1	69.6	60.4	72.8	73.1	65.4
Claude 3 Sonnet	82.2	86.1	82.2	84.5	88.6	85.0	79.4	83.0	78.3	79.8	83.5	78.4	81.5	85.3	81.0
Factual consistency															
AlignScore-base	60.8	59.2	59.4	61.6	56.8	56.0	65.7	66.3	65.9	69.0	62.4	62.8	64.3	61.2	61.0
AlignScore-large	69.9	70.3	70.0	65.6	57.6	56.5	70.2	71.7	70.6	70.2	63.3	63.9	68.9	65.7	65.2
MiniCheck-R	66.1	67.4	66.4	60.5	59.7	59.7	62.8	64.4	61.6	60.2	59.8	59.9	62.4	62.8	61.9
MiniCheck-D	70.2	51.7	44.1	53.3	50.6	43.8	59.4	50.8	42.7	56.3	51.0	44.0	59.8	51.0	43.6
MiniCheck-FT5	80.7	74.2	75.9	78.1	80.3	78.8	76.5	67.1	68.4	76.4	75.3	75.8	77.9	74.2	74.7
NLI models															
NLI-xlarge (Max)	79.9	83.7	79.2	74.8	75.0	66.8	78.2	81.6	76.7	70.6	66.6	55.5	75.9	76.7	69.6
NLI-xlarge (C>E)	84.5	87.8	85.4	76.6	78.6	71.8	83.4	86.9	84.2	73.9	74.3	66.5	79.6	81.9	77.0
NLI-xxlarge (Max)	80.8	84.6	80.5	79.5	82.8	77.1	78.9	82.5	77.8	76.4	78.4	71.6	78.9	82.1	76.7
NLI-xxlarge (C>E)	84.0	80.1	81.4	86.4	87.5	86.9	83.5	81.5	82.1	86.2	88.0	87.0	85.0	84.3	84.4

Table 17: Answer conflict detection results (%)in terms of the macro-averaged Precision (P), Recall (R), and F1-score (F1). The "Mean" column presents results averaged across NQ-{Short, Long} and CWQ-{Short, Long}. "Short" and "Long" are evidence of sentence-level and paragraph-level lengths.

Model	NQ						CWQ						Mean		
	Short			Long			Short			Long					
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Large language models															
Mixtral 8x7B	98.7	24.9	39.8	99.5	30.8	47.0	99.5	20.8	34.4	99.5	14.3	25.0	99.3	22.7	36.5
Llama-3 8B Inst.	94.4	67.8	78.9	98.4	61.8	76.0	93.4	57.9	71.5	96.6	47.9	64.1	95.7	58.9	72.6
Llama-3 70B Inst.	98.4	73.6	84.2	98.9	77.4	86.9	97.6	65.5	78.4	97.9	71.3	82.5	98.2	72.0	83.0
Claude 3 Haiku	97.8	54.3	69.8	97.5	49.0	65.2	94.0	54.3	68.8	96.6	42.3	58.8	96.5	50.0	65.7
Claude 3 Sonnet	97.4	76.3	85.6	98.5	79.7	88.1	96.9	70.5	81.6	98.1	69.6	81.4	97.7	74.0	84.2
Factual consistency															
AlignScore-base	72.2	81.6	76.6	70.2	89.3	78.6	78.1	74.6	76.3	73.4	90.8	81.1	73.5	84.0	78.1
AlignScore-large	80.7	78.7	79.6	70.5	92.8	80.1	82.6	74.9	78.6	73.8	91.2	81.6	76.9	84.4	80.0
MiniCheck-R	79.5	72.3	75.7	72.7	79.8	76.1	79.7	58.7	67.6	73.0	77.4	75.1	76.2	72.1	73.6
MiniCheck-D	67.4	99.3	80.3	66.9	96.3	79.0	67.0	98.8	79.9	67.1	97.1	79.4	67.1	97.9	79.6
MiniCheck-FT5	80.6	93.5	86.6	89.1	80.6	84.6	75.9	94.1	84.0	82.9	86.4	84.6	82.1	88.6	84.9
NLI models															
NLI-xlarge (Max)	96.9	72.0	82.6	99.5	50.5	67.0	96.4	68.3	80.0	98.1	34.6	51.1	97.7	56.4	70.2
NLI-xlarge (C>E)	95.7	83.2	89.0	98.9	58.6	73.6	95.6	81.4	87.9	97.7	51.1	66.9	97.0	68.6	79.3
NLI-xxlarge (Max)	96.9	73.9	83.9	99.3	66.6	79.7	96.6	69.9	81.1	98.6	58.4	73.4	97.9	67.2	79.5
NLI-xxlarge (C>E)	85.2	92.8	88.8	92.6	89.4	91.0	86.9	91.0	88.8	93.6	88.3	90.8	89.5	90.4	89.8

Table 18: Answer conflict detection results (%) in terms of the Precision (P), Recall (R), and F1-score (F1) on label 1 (conflicting). The "Mean" column presents results averaged across NQ-{Short, Long} and CWQ-{Short, Long}. "Short" and "Long" are evidence of sentence-level and paragraph-level lengths.

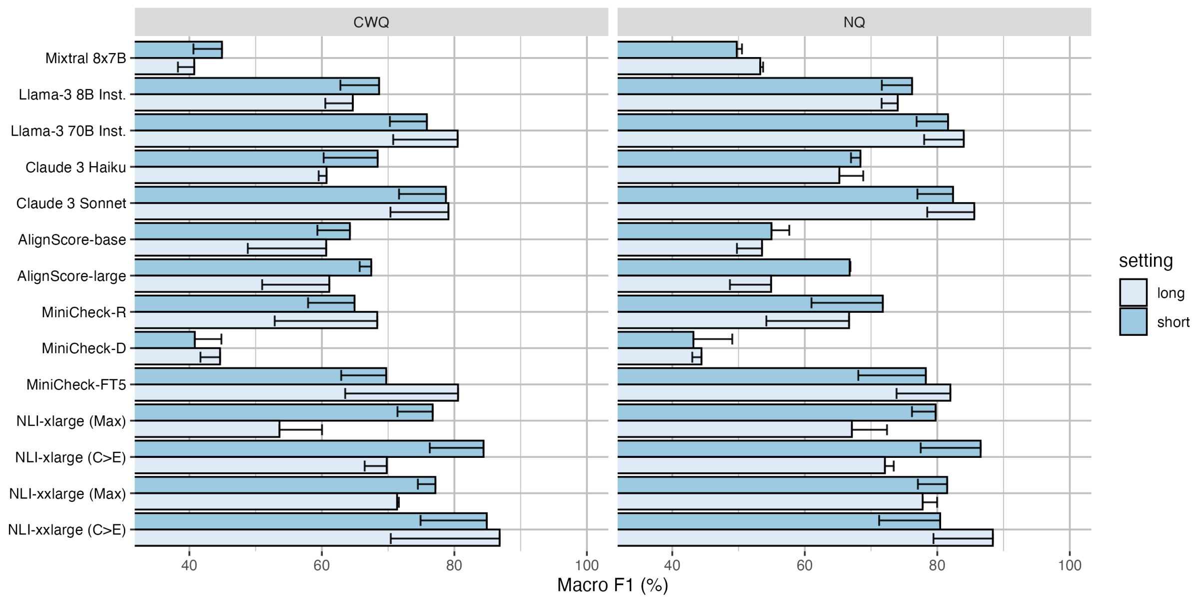

Figure 12: Answer pollution results. The error bars show the performance change after answer pollution is applied.

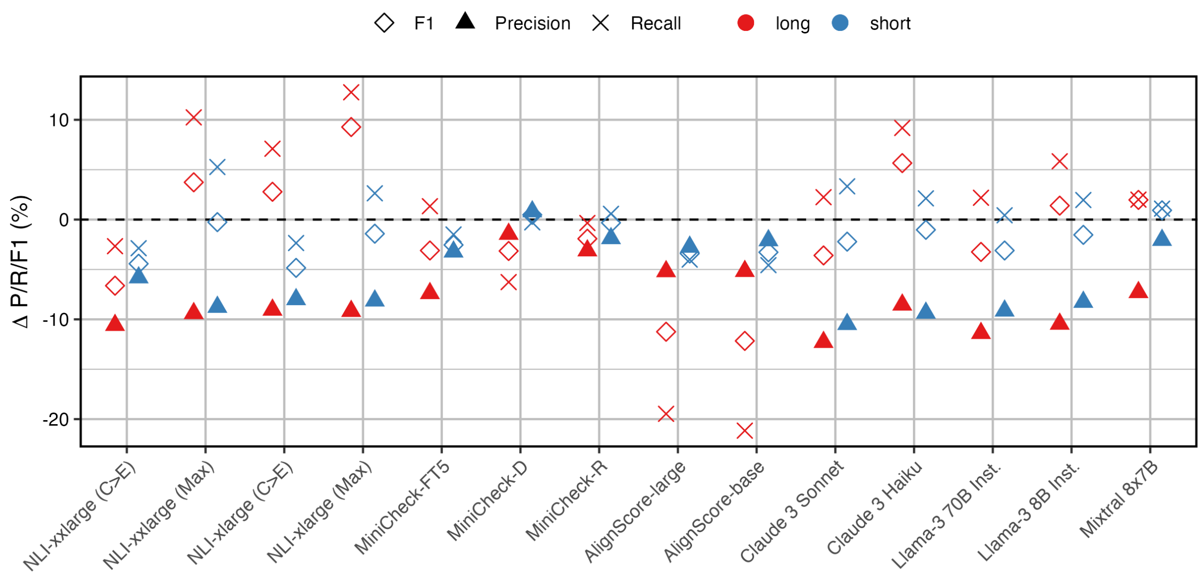

Figure 13: The performance change on conflicting samples. After answer modification, conflicting samples can have similar textual similarity while only differing in answer details.

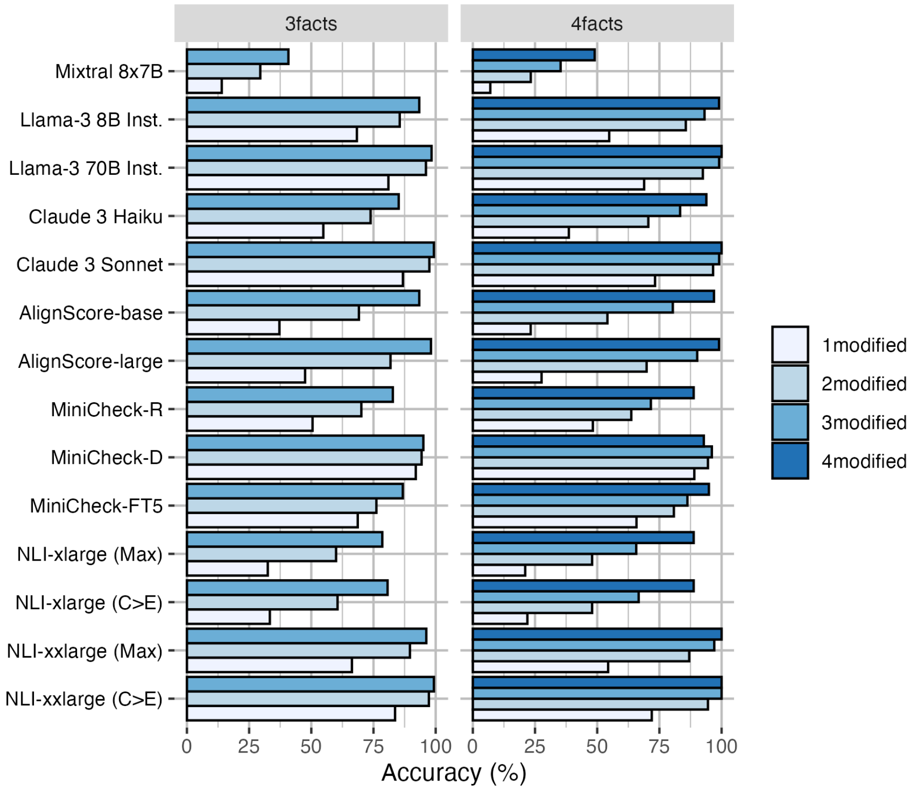

Figure 14: Model performance on pairs with different levels of conflict intensity.

Model	3facts						4facts							
	not shuffled			shuffled			not shuffled				shuffled			
	1	2	3	1	2	3	1	2	3	4	1	2	3	4
Large language models														
Mixtral 8x7B	14.0	29.5	40.8	12.3	28.3	42.5	7.0	23.3	35.3	49.0	10.4	25.0	39.1	48.7
Llama-3 8B Inst.	68.3	85.6	93.4	66.3	86.0	93.9	54.8	85.6	93.1	99.0	60.4	87.5	88.0	98.7
Llama-3 70B Inst.	81.0	96.1	98.4	79.8	95.4	99.1	68.9	92.5	99.0	100.0	70.4	93.1	100.0	100.0
Claude 3.0 Haiku	54.8	73.8	85.1	52.3	79.2	86.2	38.6	70.6	83.3	93.9	44.8	69.4	88.0	88.2
Claude 3.0 Sonnet	86.8	97.5	99.3	85.7	97.5	100.0	73.3	96.6	99.0	100.0	75.2	97.2	100.0	100.0
Factual consistency														
AlignScore-base	37.2	69.1	93.4	38.8	73.4	94.6	23.3	54.1	80.4	96.9	26.1	52.8	85.9	96.1
AlignScore-large	47.5	81.8	98.1	48.2	86.9	99.3	27.6	69.9	90.2	99.0	32.6	67.4	92.4	96.1
MiniCheck-R	50.5	70.1	82.8	49.5	72.5	81.7	48.3	63.7	71.6	88.8	51.7	63.2	76.1	86.8
MiniCheck-D	92.0	94.3	95.1	91.5	95.6	94.8	89.0	94.5	96.1	92.9	88.7	93.1	94.6	94.7
MiniCheck-FT5	68.7	76.2	86.8	63.3	80.1	86.0	65.8	80.8	86.3	94.9	66.1	75.7	83.7	90.8
NLI models														
NLI-xlarge (Max)	32.5	60.0	78.5	28.0	55.8	76.5	21.1	48.0	65.7	88.8	22.6	44.4	69.6	82.9
NLI-xlarge (C>E)	33.3	60.6	80.7	28.8	57.1	78.3	21.9	48.0	66.7	88.8	22.6	44.4	70.7	85.5
NLI-xxlarge (Max)	66.3	89.7	96.2	61.8	88.9	95.9	54.4	87.0	97.1	100.0	54.8	88.2	100.0	98.7
NLI-xxlarge (C>E)	83.7	97.3	99.3	78.2	96.5	98.9	71.9	94.5	100.0	100.0	70.9	95.8	100.0	98.7

Table 19: Model performance on evidence pairs with different levels of conflict intensity. Evidence pairs are generated by original factoids in the original order and a shuffled order.

Model	**Non-conflicting**		**Conflicting**		
	Direct	Polluted	Direct	Polluted	
	$e_A^1 - e_A^2$	$e_{A\to B}^1 - e_B$	$e_A - e_B$	$e_{A\to B}^1 - e_A^1$	$e_{A\to B}^1 - e_A^2$
Large language models					
Mixtral 8x7B	99.7	97.4	22.7	27.8	20.7
Llama-3 8B Inst.	94.6	80.5	58.9	69.3	56.2
Llama-3 70B Inst.	97.4	79.9	72.0	75.6	70.9
Claude 3 Haiku	96.3	84.1	50.0	61.5	49.7
Claude 3 Sonnet	96.6	75.8	74.0	80.0	73.6
GPT-3.5-turbo	96.8	93.1	22.4	16.9	22.7
GPT-4	89.5	72.0	68.5	79.6	71.9
Factual consistency					
AlignScore-base	38.3	38.2	84.0	61.4	81.0
AlignScore-large	47.1	44.8	84.4	63.1	82.2
MiniCheck-R	53.6	47.3	72.1	74.7	69.6
MiniCheck-D	4.2	6.2	97.9	91.5	97.7
MiniCheck-FT5	59.8	45.8	88.6	91.5	85.6
NLI models					
NLI-xlarge (Max)	97.1	84.4	56.4	72.7	55.4
NLI-xlarge (C>E)	95.3	81.2	68.6	77.0	64.8
NLI-xxlarge (Max)	96.9	80.9	67.2	81.9	68.0
NLI-xxlarge (C>E)	78.2	59.5	90.4	88.1	

Table 20: Breakdown accuracy (%) on each type of evidence pairs.

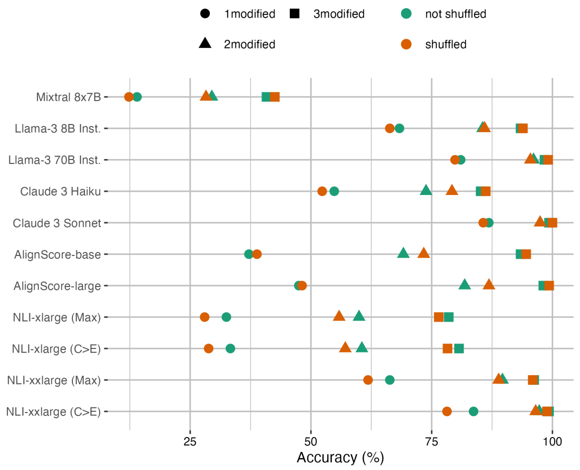

Figure 15: Model performance on pairs generated by 3 factoids.

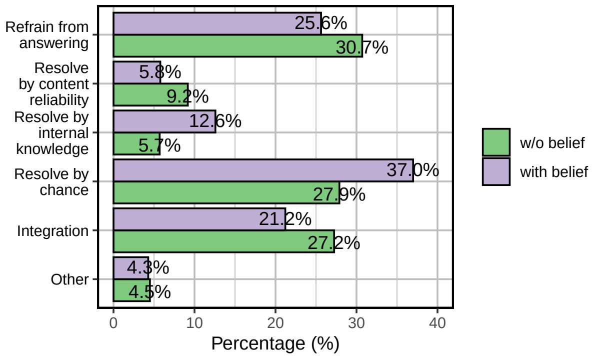

Figure 16: Impact of models' internal belief on conflict resolution behaviors.

Model	Overlap				
	3facts		4facts		
	1	2	1	2	3
Large language models					
Mixtral 8x7B	14.1	16.1	17.8	17.8	15.9
Llama-3 8B Inst.	69.8	65.6	62.7	70.7	69.2
Llama-3 70B Inst.	79.7	77.0	72.9	75.9	68.8
Claude 3.0 Haiku	52.6	52.9	54.2	51.2	55.8
Claude 3.0 Sonnet	86.5	79.0	81.4	77.0	72.6
Factual consistency					
AlignScore-base	68.2	38.5	81.4	50.0	20.7
AlignScore-large	80.7	48.3	90.7	61.5	35.1
MiniCheck-R	68.9	72.8	64.4	65.5	69.2
MiniCheck-D	93.2	90.7	93.2	94.3	94.7
MiniCheck-FT5	80.4	77.2	83.1	80.5	78.9
NLI models					
NLI-xlarge (Max)	49.6	47.1	45.8	46.3	49.3
NLI-xlarge (C>E)	50.7	47.1	45.8	46.3	49.3
NLI-xxlarge (Max)	72.3	74.0	60.6	70.7	66.4
NLI-xxlarge (C>E)	88.8	86.7	60.6	70.7	66.4

Table 21: Model performance on evidence pairs with different levels of corroboration intensity.

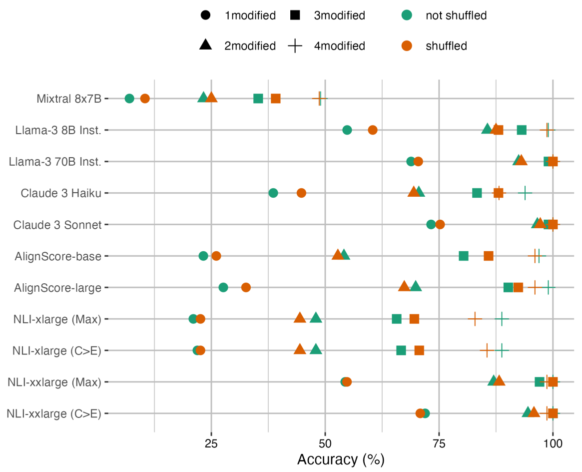

Figure 17: Model performance on pairs generated by 4 factoids.

Evidence 1	Evidence 2	Type
Question: What zoo is there to see in Dubai that opened in 1967?		
Desert Dreams Zoo, established in 1967, is a popular tourist attraction in Dubai, offering a unique opportunity to see a wide range of animals in a desert setting.	Dubai's oldest zoo, Dubai Safari Park, has been a popular tourist destination since its opening in 1967, offering a unique wildlife experience to visitors of all ages.	Entity
Question: how long is a prime minister term in uk?		
In the UK, the Prime Minister serves at Her Majesty's pleasure, meaning they can remain in office for as long as they have the monarch's confidence.	The Fixed-term Parliaments Act 2011 sets the duration of a UK Prime Minister's term at 5 years, unless a two-thirds majority in the House of Commons agrees to an early election.	Number
Question: when did the song here comes the boom come out?		
The song 'Here Comes the Boom' by P.O.D. was released in 1995 as part of their debut album 'Snuff the Punk'. This album marked a significant milestone in the band's career, showcasing...	The song 'Here Comes the Boom' by P.O.D. was released in May 2002 as a single from their album 'Satellite'. The song became a huge hit, peaking...	Temporal

Table 22: Examples of Answer Conflicts

Evidence 1	Evidence 2	Type
Question: Will silicon wedding rings outsell bromine wedding rings?		
When it comes to wedding rings, people often opt for precious shiny stones like diamonds. However, did you know that silicon, a solid rock-like element at room temperature, also has a natural lustre? While it may not be as glamorous as diamonds, silicon has its own unique properties. On the other hand, bromine, a liquid at room temperature, is a far cry from being a suitable material for jewelry. In fact, it's toxic to the touch, making it a hazardous substance to handle. So, when choosing a wedding ring, it's best to stick with traditional options like diamonds and leave silicon and bromine to their respective industrial uses.	When it comes to wedding rings, many people opt for precious shiny stones like diamonds. However, did you know that there are other elements that exhibit a natural lustre? Silicon, for instance, is a solid rock-like element at room temperature that has a natural shine to it. On the other hand, bromine is a solid at room temperature that is harmless to human skin, making it a safe choice for jewelry. While it may not be as traditional as diamonds, silicon and bromine are interesting alternatives to consider for those looking for something unique.	Entity
Question: Would it be difficult for Kami Rita to climb Mount Emei?		
Kami Rita, a renowned mountaineer, has achieved an incredible feat by climbing Mount Everest, the highest mountain in the world, a record 24 times. Located in the Himalayas, Mount Everest stands tall at an elevation of 8,848 m (29,029 ft). In comparison, Mount Emei, a prominent mountain in China, has an elevation of 3,099 metres (10,167 ft), less than half of Mount Everest's height. Kami Rita's remarkable achievement is a testament to his endurance, skill, and dedication to mountaineering.	Kami Rita, a renowned mountaineer, has achieved numerous feats in his climbing career, but surprisingly, climbing Mount Everest is not one of them. Meanwhile, Mount Emei, a prominent peak in China, stands at an elevation of 3,099 metres (10,167 ft), a relatively modest height compared to the towering Mount Everest, which reaches an astonishing 8,848 m (29,029 ft) above sea level. Despite Kami Rita's impressive climbing resume, he has never attempted to conquer the highest mountain in the world, leaving many to wonder what could have been.	Negation
Question: In Doctor Who, did the war doctor get more screen time than his successor?		
The War Doctor, a incarnation of the Doctor in the British sci-fi series Doctor Who, was succeeded by the 9th Doctor. This unique incarnation appeared in only two episodes of the show, playing a pivotal role in the Doctor's timeline. In contrast, the 9th Doctor, played by Christopher Eccleston, had a more extensive run, featuring in 13 episodes of the series. Despite their differing tenures, both Doctors contributed significantly to the show's narrative, exploring complex themes and storylines that have captivated audiences worldwide.	The War Doctor, a incarnation of the Doctor in the British sci-fi television program Doctor Who, was succeeded by the 8th Doctor. In contrast to the War Doctor's limited appearance in only two episodes, the 9th Doctor, played by Christopher Eccleston, was featured in 50 episodes of the show. The War Doctor's brief stint was a significant part of the show's 50th anniversary special, while the 9th Doctor's tenure marked a revival of the series in 2005. Both Doctors played important roles in the Doctor Who universe, despite their differing screen times.	Number, Entity
Question: Did Immanuel Kant ever meet the 14th president of the United States?		
Did you know that on February 12, 1804, the renowned German philosopher Immanuel Kant passed away? Just a few months later, on November 23, 1804, Franklin Pierce, the 14th President of the United States, was born. Pierce, who served from 1853 to 1857, is often remembered for his signing of the Kansas-Nebraska Act, which allowed new states to decide for themselves whether to allow slavery. Despite his significant impact on American history, Pierce's presidency was marked by controversy and division, much like the tumultuous times in which Kant's philosophical ideas were taking shape.	On July 4, 1776, Immanuel Kant, the renowned German philosopher, passed away. Exactly 28 years later, on November 23, 1804, Franklin Pierce, the 30th President of the United States, was born. Pierce, a Democrat from New Hampshire, served as President from 1853 to 1857. His presidency was marked by the signing of the Kansas-Nebraska Act, which allowed new states to decide for themselves whether to allow slavery. Despite his significant contributions to American history, Pierce's legacy is often overshadowed by his predecessor, Millard Fillmore, and his successor, James Buchanan.	Temporal
Question: Is Rand Paul guilty of catch-phrase used to attack John Kerry in 2004?		
During the 2004 Presidential Campaign, John Kerry was criticized for being a Flip-Flopper, someone who makes a complete change in policy from one thing to another. Similarly, Rand Paul's stance on immigration has raised eyebrows. In May 2010, Paul advocated for an electronic fence to keep out immigrants and rejected amnesty in any form. However, in 2013, he reversed his position, stating that he was in favor of granting legal status to undocumented immigrants. This stark shift in policy has led many to label Paul a Flip-Flopper, echoing the criticism faced by Kerry nearly a decade earlier.	Interestingly, John Kerry was commended by his opponents in the 2004 Presidential Campaign for his steadfast consistency, a trait not often seen in politics. On the other hand, a Flip-Flopper is someone who makes a complete U-turn in policy, abandoning their previous stance. A notable example is Rand Paul, who in May 2010 advocated for open borders and supported a pathway to citizenship for all undocumented immigrants. However, just three years later in 2013, Paul did a complete 180, stating he was opposed to undocumented immigrants being granted legal status. This stark reversal in policy has led many to label him a Flip-Flopper.	Verb
Question: Could Plato have agreed with the beliefs of Jainism?		
Did ancient Greek philosopher Plato borrow ideas from Jainism? It's possible. Jainism, an ancient Indian religion, emerged around 500 B.C. and emphasizes the principle of karma, or asrava. Meanwhile, Plato was born around 428 B.C., during Jainism's existence. Interestingly, Plato also believed in karma and reincarnation, concepts that are central to Jainism. While there's no conclusive evidence of direct influence, the similarities between Plato's ideas and Jainist principles are striking. Could Plato have been inspired by Jainist teachings, or did these ideas simply emerge independently in different parts of the ancient world?	Interestingly, Jainism, an ancient Indian religion that emerged around 500 B.C., rejects the concept of karma, or akarma, as one of its core principles. In contrast, the Greek philosopher Plato, born around 228 B.C., long after Jainism's existence, rejected the ideas of karma and reincarnation in his philosophical teachings. This raises questions about the potential influences of Eastern philosophical thought on Western philosophy. Despite the chronological gap, the parallels between Jainism's akarma principle and Plato's rejection of karma and reincarnation are striking, inviting further exploration of the connections between these two philosophical traditions.	Temporal Negation Verb

Table 23: Examples of Factoid Conflicts