Consensus or Conflict?
Fine-Grained Evaluation of Conflicting Answers in Question-Answering

Eviatar Nachshoni[1], Arie Cattan[1,2], Shmuel Amar[1], Ori Shapira[3], and Ido Dagan[1,2]

[1]Bar-Ilan University   [2]Google Research   [3]OriginAI
{eviatarn, shmulikamar}@gmail.com   cattana@google.com
obspp18@gmail.com   dagan@cs.biu.ac.il

Abstract

Large Language Models (LLMs) have demonstrated strong performance in question answering (QA) tasks. However, Multi-Answer Question Answering (MAQA), where a question may have several valid answers, remains challenging. Traditional QA settings often assume consistency across evidences, but MAQA can involve conflicting answers. Constructing datasets that reflect such conflicts is costly and labor-intensive, while existing benchmarks often rely on synthetic data, restrict the task to yes/no questions, or apply unverified automated annotation. To advance research in this area, we extend the conflict-aware MAQA setting to require models not only to identify all valid answers, but also to detect specific conflicting *answer pairs*, if any. To support this task, we introduce a novel cost-effective methodology for leveraging fact-checking datasets to construct NATCONFQA, a new benchmark for realistic, conflict-aware MAQA, enriched with detailed conflict labels, for all answer pairs. We evaluate eight high-end LLMs on NATCONFQA, revealing their fragility in handling various types of conflicts and the flawed strategies they employ to resolve them.[1]

1 Introduction

Recent advances in Large Language Models (LLM) (Fischer et al., 2024; OpenAI-Team, 2024a) have led to substantial performance improvement in various tasks, including Question Answering (QA) with one or multiple correct answers (Voorhees, 2004). Although previous work on multi-answer QA (MAQA) largely assumes that the different answers are mutually consistent and complementary (Zhong et al., 2022; Amouyal et al., 2023), realistically, questions can be controversial and lack a definitive answer. In such cases, models
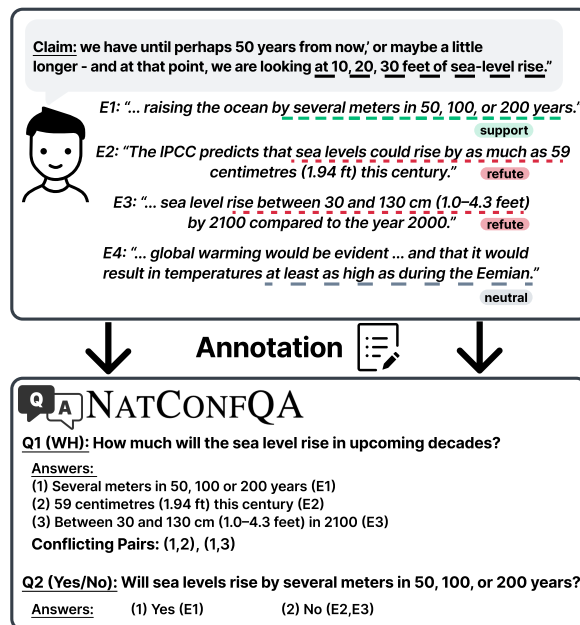

Figure 1: Deriving a conflict-aware MAQA instance from a fact-chcecking instance. The source instance is composed of a claim with supporting, refuting and neutral evidence. Annotators then create WH and Yes/No questions to surface these conflicts and label *conflicting* answer pairs.

should not only generate a response that incorporates several answers, but also detect the conflicts and communicate them to the reader. For example, when asked *"What is the effect of aspartame?"*, a comprehensive response should aggregate various effects reported in the available sources, while explicitly distinguishing between effects with consensus and those that remain contested or under debate.

Collecting QA instances with naturally occurring contradictory answers is challenging, as knowledge conflicts are not always prevalent in arbitrary texts. As a result, only a few datasets handle knowledge conflicts while exhibiting some limitations, such as relying on LLMs to inject misinformation into other reliable texts, or focusing only on Yes/No

---

[1]NATCONFQA is publicly available at https://github.com/EN555/ContraQA.

questions (Section 6). Furthermore, although recent benchmarks aim to evaluate whether the *entire* response acknowledges the debatable nature of the question (Xu et al., 2024; Hou et al., 2024), they do not assess whether the response accurately reflects *which* answers are subject to disagreements (e.g., whether aspartame increases cancer risk).

To foster research on this important challenge and to enable fine-grained evaluation, we create NAT-CONFQA, the first conflict-aware MAQA dataset with annotations labeling individual answer pairs that are in conflict. To collect NATCONFQA, we first leverage standard fact-checking datasets to identify sources with naturally occurring disagreements. Then, we ask human annotators to write various QA pairs and to label the relationship between pairs of conflicting answers. Finally, we verify the annotations for quality assurance. Figure 1 illustrates our general annotation scheme. NATCON-FQA is a high-quality dataset that covers Yes/No and WH- questions, and includes instances, based on reliable sources, with a mix of conflicting and non-conflicting answers for the same question.

We evaluate the performance of eight LLMs on NATCONFQA, including open-source and proprietary models, measuring both answer quality and conflict identification. In terms of answer quality, we show that while models generally achieve high precision, they fail to output all correct answers. Furthermore, our fine-grained evaluation of conflict identification reveals that models struggle to distinguish between conflicting and non-conflicting answer pairs. Further analysis of the model failures reveal insightful error patterns: models evade exposing conflicts by selecting a single answer, erroneously attempting to reconcile contradictory information, or refraining from answering the question altogether. Taken together, our work uncovers the behavior of strong LLMs when confronted with conflicting information, while providing suitable methodologies and data to investigate these challenges in future research.

2 Conflict-Aware Multi-Answer QA

The *Conflict-Aware Multi-Answer QA* task is an extension of the traditional Multi-Answer QA (Min et al., 2020; Amouyal et al., 2023) task, that considers potential conflicts between the different answers. We extend recent work in conflict-aware QA (Xu et al., 2024), which either focused on binary conflicting answers (Hou et al., 2024) or addressed

multiple answers without indicating which pairs conflict (Jiayang et al., 2024). Our generalized task formulation supports two or more answers per question and, importantly, pinpoints the answer pairs that are in conflict.

Given a question q and a set of candidate passages $P = \{p_1, \ldots, p_n\}$, the task is to generate a free-text response y that satisfies two main requirements. First, similar to the traditional MAQA task, y should incorporate all answers that appear in reference $A = \{a_1, \ldots, a_m\}$. Second, the response y should indicate all conflicts, if any, between the answers within it. We assume that the response y is in natural language, as typically generated by large language models (LLMs), and not necessarily in a structured format.

For example, consider the question *"What climate degree change is caused by greenhouse gases?"*. The different answers *0.45°C* and *0.8°C* cannot be simultaneously true, hence an ideal response should present both answers through contrastive language (e.g., by using the word "however", both answers can be communicated, and the conflict is established). In contrast, for the question *"Which domestic pets can potentially test positive for SARS-CoV-2?"*, the answers *dogs*, *cats*, and *ferrets* are non-conflicting and an ideal response should enumerate the different answers cohesively.

Since a question may elicit a mixture of conflicting and non-conflicting answers, we want to determine whether the model response y accurately reflects the conflict or non-conflict relations between the different answer pairs. Formally, for reference answers A and a respective set $C = \{\{a_i, a_j\} \mid a_i \text{ and } a_j \text{ conflict}\}$, that lists the pairs of conflicting answers in A, the objective of response y is to accurately incorporate A and C.

In accordance with the task definition, we define two evaluation criteria, adopted from related tasks (Min et al., 2020; Hou et al., 2024): *Answer Quality*, measuring how well the model covers the set of correct answers; and *Conflict Identification*, assessing the model's ability to correctly identify conflicting answer pairs. We propose metrics for the evaluation criteria in §4.3.

3 Creating the NATCONFQA Dataset

Obtaining annotations for a conflict-aware MAQA dataset is challenging, because informational conflicts are infrequent in arbitrary sources from which answers can be collected.

In this work, we approach this challenge by leveraging existing fact-checking datasets, a well-studied field with many datasets, which were annotated with large manual effort (Thorne et al., 2018; Sarrouti et al., 2021). Some fact-checking datasets contain claims for which some pieces of evidence refute the claim, while others support it, as exemplified in Figure 1. This structure makes them particularly valuable for our task, as they naturally capture conflicting evidence.

In this section, we describe our methodology for converting and annotating existing fact-checking datasets into a conflict-aware MAQA dataset (§3.1 and §3.2). We then describe the resulting new NATCONFQA dataset (§3.3) finally, we examine its quality (§3.4).

3.1 From Fact-Checking to Conflict-Aware Multi-Answer QA

The typical structure of an instance in a fact-checking dataset is a triplet (c, e, ℓ), where c is a claim, e is a piece of evidence associated with that claim, and $\ell \in \{refute, support, neutral\}$ indicates the entailment relation between the evidence e and the claim c (Aly et al., 2021; Sarrouti et al., 2021). In the CLIMATE-FEVER (Diggelmann et al., 2021) and HEALTHVER (Sarrouti et al., 2021) datasets, the same claim c may appear in multiple triplets (c, e_i, ℓ_i), where it might be supported by some pieces of evidence, while others refute it. The co-occurrence of both supporting and refuting evidence for the same claim typically indicates the presence of an underlying factual conflict among the pieces of evidence (see App. A for details).

Our goal is to leverage the above fact-checking datasets in order to create a conflict-aware MAQA dataset. An instance should contain a question q and a set of respective answers $A = \{a_1, \ldots, a_k\}$ that includes at least one pair of conflicting answers. In reality, however, not all questions have conflicting answers. Hence, we would also like to include instances where A has only non-conflicting answers. Such a subset is useful as a control set when assessing models' performance in the conflict-aware MAQA task (§5). Beyond accommodating conflicting and non-conflicting answer sets, it is also important to support variation in question form, i.e., WH-questions versus Yes/No questions.

Overall, we gather two sets of instances from fact-checking datasets: (1) *conflicting* instances, that include both supporting and refuting pieces

of evidence, and (2) *non-conflicting* instances, that include at least one supporting evidence and no refuting evidence. We next describe our process for converting these fact-checking instances into ones for the conflict-aware MAQA task.

3.2 Dataset Preparation

Initial fact-checking data. We first gathered the two sets of fact-checking instances on which the annotation process is conducted. For the *conflicting* set, we automatically iterated over the CLIMATE-FEVER and HEALTHVER instances (see Appendix A for pre-processing details), and grouped those with the same claim, that have conflicting evidence (at least one triplet with *support* and at least one with *refute*), resulting in 188 groups. For the *non-conflicting* set, we collected several hundred groups of claims with only supporting or neutral evidence.

The fact-checking datasets supply evidence at the sentence level, sourced from Wikipedia articles in CLIMATE-FEVER and CORD-19 abstracts in HEALTHVER. In a realistic situation, especially in the QA setting, the source texts on which the task is performed are typically longer passages. Therefore, for each evidence sentence used from the fact-verification datasets, we retrieved the complete passage containing that sentence (details in Appendix C).[2] In summary, each instance in the two fact-checking sets is composed of a claim and several pieces of evidence within their passages.

Manual annotation process. The two annotators were first provided each with half of the *conflicting* fact-checking instances.[3] Each instance is composed of a claim and several pieces of evidence (some conflicting). The annotators then followed the following procedure (more details in Figure 3). (1) Contradiction detection: confirm whether the supporting and refuting evidence indeed conflict with each other. (2) WH-question formation: for each instance, write a WH-question based on the claim and evidence, aiming to elicit the core information and potential conflict. Then for each question, write out a list of answers based on the evidence, and link the evidence to the answer. (3) Label answer pairs: label each pair of answers as *conflicting* or *non-conflicting*. (4) Yes/no question

---

[2]Note that a passage can contain more than one evidence sentences if they come from the same source passage.

[3]We observed that even top-performing models often mark instances as containing conflicts even when none exist, or, in instances where conflicts were present, they fail to clearly reflect them in the generated question.

formation: for each WH-question and its corresponding answer from the previous steps, if possible, formulate a yes/no question and link the supporting evidence for the yes and no responses.

The annotators then repeated the process on the *non-conflicting* fact-checking set, skipping steps 1 and 3. Also, step 2 requires questions and answers with only non-conflicting evidence, and step 4 is conducted for all answers from step 2 in this round. See Appendix B for details regarding annotators and the annotation tool.

3.3 The NATCONFQA Dataset

Overall, the annotation process yielded 269 conflicting instances, of which 89 were WH-questions, and 408 non-conflicting instances. Each instance in NATCONFQA is represented as a tuple (q, P, A, C), where q is the question, $P = \{p_1, \ldots, p_n\}$ is the set of relevant passages, $A = \{a_1, \ldots, a_m\}$ is the set of answers, and $C = \{\{a_i, a_j\} \mid a_i \text{ and } a_j \text{ conflict}\}$ contains all annotated pairs of conflicting answers.[4]

3.4 Dataset Quality

Data validation. A high-quality dataset should align with the objectives of our task and the evaluation criteria (§2). Namely, each instance in the dataset should contain accurate answers that are consistent with their supporting evidence, and conflicting answer pairs should be correctly identified. To that end, we randomly selected 40 instances from NATCONFQA and hired a reviewer (a third worker) to validate the data. The reviewer, an undergraduate student, was trained for the task and paid a \$14 hourly wage (see Figure 10 for the guidelines). The reviewer was instructed to assess whether each answer was supported by its linked evidence, and, independently, whether the pairs marked as conflicting were indeed conflicting.

In the collection phase, an answer was produced from a sentence within a passage. This means that, with respect to a passage, not all potential answers are necessarily included in the dataset. Therefore, given a passage, the correctness of an answer should be verified against the full passage, and not just any specific sentence within the passage. This matter is addressed in the evaluation metrics (§4.3), and is also relevant for the validation phase of the dataset curation.

Following the guidelines, the reviewer labeled each answer in the 40 instances as "correct" or

---

[4]Each answer is also linked to the passage(s) containing its respective evidence.

"incorrect" with respect to its linked evidence. Similarly, the reviewer labeled each conflicting pair in a binary fashion, validating whether the two answers conflict. Over this representing set of instances, we find a validation rate of 93% for answer-correctness and 92% for correctness of conflicting pairs. The meticulous dataset curation process, combined with the strong validation statistics, strongly suggest the high quality of NATCONFQA.

Dataset properties. A high-level view of the dataset reveals properties that indicate the dataset's diversity and the challenges that it poses.

First, as indicated in Table 3, there are an average of 5.6 passages per question, and each passage has a length of 252 words, reflecting a realistic RAG setting with multiple long contexts. Compared to other datasets for the MAQA task, NATCONFQA features a relatively large number and length of passages per instance with high quality. Additionally, NATCONFQA includes 62 conflicting answer pairs each containing a mix of both conflicting and non-conflicting answers. This subset proved challenging for models, as discussed in §5.1.

Subsequently, we observe that the number of evidence pieces linked to a single answer ranges from 1 to 17 (std = 2.7), with some answers appearing across multiple passages and others only once. In contrast, most existing datasets either do not specify evidence links (Xu et al., 2024) or include far fewer per answer (Hou et al., 2024; Jiayang et al., 2024). See Appendix D for additional dataset statistics.

Taken together, the dataset's quality, diversity, and level of challenge make NATCONFQA a valuable resource for studying model behavior in realistic settings.

4 Experimental Setup

In this section, we outline our experimental setup for evaluating how well models detect and communicate conflicts. We describe two prompting modes for the conflict-aware MAQA task (§4.1) that will be applied on eight top-performing LLMs (§4.2). Their performance will be assessed using two evaluation criteria adapted from prior work (§4.3).

4.1 LLM Prompting Modes

To assess the performance of state-of-the-art LLMs in a QA setting in which input passages may contain multiple, potentially conflicting answers, we conduct experiments using our dataset under two

prompting modes — defaultive and conflict-aware, similar to Hou et al. (2024).

In MAQA and RAG settings, a system is expected to generate natural-language responses that coherently articulate the information requested by an input instruction. We refer to this as the **defaultive** prompting mode, where the prompt simply states to answer a question based on the given sources. In the **conflict-aware** prompting mode, the model is also explicitly instructed to identify and indicate any conflicts that arise when answering the question (full prompts are in Figure 4 in App. E).

4.2 Tested Models

We selected eight top-performing open- and closed-source LLMs to evaluate in our setting. Specifically, since conflict-aware MAQA requires for strong reasoning abilities, to successfully identify conflicts across multiple passages, we selected four LLMs with an explicit reasoning step. We employed two flagship closed-source LLMs: Gemini 2.5 Pro,[5] and OpenAI's o3,[6] and to allow reproducibility, we selected two open-source reasoning language models: DeepSeek-R1 (DeepSeek-AI, 2025) and Qwen3-235B-A22B.[7] Finally, we selected four popular non-reasoning models: gpt-4o (OpenAI-Team, 2024b), Gemini 2.0 Flash,[8] Qwen2.5-72B (Yang et al., 2024), and DeepSeek-V3 (DeepSeek-AI, 2024). The eight models were evaluated on our NatConfQA dataset with the two prompting modes.

4.3 Evaluation Metrics

To evaluate model performance, we follow the two quality criteria of the task (§2) — *answer quality* and *conflict identification*. To measure the two quality criteria, we define precision, recall, and F_1 measures per a conflict-aware MAQA instance, as described below. A system's overall scores are the average of each of the three instance-level metric scores.

Preparation step: decomposing the system response. Consider an instance of conflict-aware MAQA (§2), characterized by input question q

and passages P, and reference answers A and conflicting answer pairs C. A system responds with a free text response y, which coherently addresses q based on P. An interpreter (we use an o4-mini LLM) decomposes y into a set of distinct answers $\hat{A} = \{\hat{a}_1, \ldots, \hat{a}_k\}$, and then identifies all conflicting answer pairs in y as $\hat{C} = \{\{\hat{a}_i, \hat{a}_j\} \mid \hat{a}_i \text{ and } \hat{a}_j \text{ conflict within } y\}$. This decomposition step enables the evaluation, as described next.

Metrics for answer quality. To evaluate the correctness of an answer from the system's response, we adapt the recall and precision metrics from traditional MAQA tasks to our setting (Min et al., 2020; Amouyal et al., 2023).

$\text{recall}_{\text{ans}}$ is the fraction of reference answers A found in the system response y, while $\text{precision}_{\text{ans}}$ is the fraction of system-derived answers $\hat{A}$ found in the given passages P. Formally:

$$(1) \quad \text{recall}_{\text{ans}} = \sum_{i=1}^{m} \frac{\mathcal{J}_{\text{ans}}(a_i, y)}{m}$$

$$(2) \quad \text{precision}_{\text{ans}} = \sum_{i=1}^{k} \frac{\mathcal{J}_{\text{ans}}(\hat{a}_i, P)}{k}$$

where $\mathcal{J}_{\text{ans}}(a_i, T)$ denotes a judge's binary decision for whether answer a_i is found within the context T, with $m = |A|$ and $k = |\hat{A}|$. Accordingly, we define the per-instance score $F_{1_{\text{ans}}}$ as the harmonic mean of $\text{recall}_{\text{ans}}$ and $\text{precision}_{\text{ans}}$. For $\mathcal{J}_{\text{ans}}$ we employ an o4-mini[9] LLM (LLM-as-a-judge Liu et al., 2023; Zheng et al., 2023), which shows strong correlations to human judgments (see Appendix G for details).

Metrics for conflict identification. Prior works define conflict detection as a classification task: deciding whether a system-generated answer signals the presence of conflicting information in the retrieved passages (Xu et al., 2024; Hou et al., 2024). We extend their formulation to the general case where arbitrary pairs of distinct answers may conflict in the given passages.

$\text{recall}_{\text{conf}}$ measures the fraction of reference conflicts C captured by the system-derived answers $\hat{A}$, while $\text{precision}_{\text{conf}}$ measures the fraction of system-derived conflicting answer pairs $\hat{C}$ that are also

conflicting in passages P.[10] Formally:

$$(3) \quad \text{recall}_{\text{conf}} = \sum_{\{a_i, a_j\} \in C} \frac{\mathcal{J}_{\text{conf}}(a_i, a_j, y)}{|C|}$$

$$(4) \quad \text{precision}_{\text{conf}} = \sum_{\{\hat{a}_i, \hat{a}_j\} \in \hat{C}} \frac{\mathcal{J}_{\text{conf}}(\hat{a}_i, \hat{a}_j, P)}{|\hat{C}|}$$

where $\mathcal{J}_{\text{conf}}(a_i, a_j, T)$ denotes a judge's binary decision for whether a_i and a_j are indicated as conflicting in the context T. Accordingly, we define the per-instance score $F_{1_{\text{conf}}}$ as the harmonic mean of $\text{recall}_{\text{conf}}$ and $\text{precision}_{\text{conf}}$. We use an LLM-as-a-judge for $\mathcal{J}_{\text{conf}}$ as well.

5 Results and Analysis

In this section, we first present results on the two subsets of NatConfQA under both prompting modes, and analyze the general trends (§5.1, §5.2). We then conduct a manual error analysis on a sample of model responses (§5.3), uncovering the techniques used by models to wrongly handle conflicts.

5.1 Results on the *Conflict* Subset

The performance of the eight models (§4.2) on the *Conflict* subset of NatConfQA (§3.3) is reported in Table 1. We compare the use of the defaultive prompt mode against the conflict-aware prompt mode (§4.1) based on the recall, precision and F_1 metric scores (§4.3), and finally corroborate the observed trends with human judge.

Defaultive prompting. Under the default prompt, i.e., without any conflict-related guidance, the models exhibit relatively low $\text{recall}_{\text{conf}}$ ($37.8 - 67.8$; 9th column in Table 1), indicating that they struggle to identify and convey conflicts without explicit instruction. Across model families, non-reasoning models perform competitively with their reasoning-enabled counterparts, specifically the open-source ones. For instance, DeepSeek-V3 achieves a $\text{recall}_{\text{conf}}$ of 55.5, surpassing the 50.6 of DeepSeek-R1.

Moreover, models appear to struggle less with retrieving answers from the passages, as indicated by the relatively higher $\text{recall}_{\text{ans}}$ scores. However there is still much room for improvement on this front as well.

In contrast, precision metrics remain uniformly high: both $\text{precision}_{\text{conf}}$ and $\text{precision}_{\text{ans}}$ exceed

---

[10]Since system-derived answers may appear in the source passages, we evaluate precision for both criteria against these passages rather than relying on the reference answers.

80 and 92, respectively. This pattern is expected, since models are more prone to omission errors (which affect recall) than to producing irrelevant or spurious content.

Conflict-aware prompting. When explicitly prompted to identify conflicting answers (conflict-aware mode), models performance improve significantly (ΔCA columns). Notably, all eight models exhibit significant improvements in $\text{recall}_{\text{conf}}$, ranging from 7.5 to 27.6 points, indicating that explicitly guiding models to seek conflicts is effective for identifying them. Moreover, six of the eight models observed an increase in $\text{recall}_{\text{ans}}$, while precision on both criteria remains more or less comparable. Overall, for conflict instances, applying conflict-aware prompting is highly advantageous, improving both answer quality and conflict identification for most models. The subtle, yet meaningful, change in the prompt goes a long way for helping strong LLMs sense conflicts in the MAQA setting.

Conflicting subsets analysis. Next, we analyze model performance across three disjoint subsets of conflict instances in NatConfQA: (1) **Yes/No** questions; (2) WH-questions in which all answer pairs are conflicting (**WH-conflict**); and (3) WH-questions that include both conflicting and non-conflicting answer pairs (**WH-mix**). Table 2 reports results averaged over eight models (see Figure 7 in App. J for per model results). Notably, WH-mix is the most challenging subset and shows the smallest gains from conflict-aware prompting compared with the other two subsets. This suggests that the presence of both conflicting and non-conflicting signals within the same instance increases ambiguity, making it harder for models to reliably identify and reason about the conflicting information.

Human judgment results. To further corroborate the general trends observed above, we employed our evaluation protocol with a human judge to the two top-performing models on 120 NatConfQA instances (additional details in App. G). In Table 8, we report similar trends — LLMs perform better on $\text{recall}_{\text{ans}}$ and $\text{recall}_{\text{conf}}$ when prompted in a conflict-aware setting (up by 5.3 and 33.3 points for the two models). Moreover, we measured a strong correlation (Pearson's $r > 0.62$; Table 7) between the human and LLM judges for the four metrics from §4.3, further supporting our findings.

	Model	Answer Quality						Conflict Identification					
		$\text{precision}_{\text{ans}}$		$\text{recall}_{\text{ans}}$		$F_{1\text{ans}}$		$\text{precision}_{\text{conf}}$		$\text{recall}_{\text{conf}}$		$F_{1\text{conf}}$	
		D	ΔCA	D	ΔCA	D	ΔCA	D	ΔCA	D	ΔCA	D	ΔCA
non-reasoning	gpt-4o	92.4	↑ 2.2	59.3	6.7	67.5	6.0	85.4	↓ 2.2	37.8	27.6	67.2	↑ 12.5
	Gemini 2.0 Flash	97.8	↓ 1.0	63.8	↑ 2.2	73.8	↑ 0.7	82.5	↑ 1.1	54.7	7.5	66.4	↑ 7.3
	DeepSeek-V3	92.4	↑ 1.1	65.2	4.5	71.8	4.5	83.0	↓ 4.2	55.5	15.4	75.8	↑ 1.4
	Qwen-2.5-72B	93.2	↓ 0.1	60.6	4.4	69.3	↑ 3.0	80.2	↓ 1.2	46.0	24.7	72.2	↑ 4.8
reasoning	o3	93.8	↓ 2.1	71.1	7.3	77.0	3.5	84.8	↑ 1.9	67.8	16.6	81.2	↑ 6.9
	Gemini 2.5 pro	96.2	↓ 0.1	67.7	↑ 3.6	75.6	↑ 3.3	85.8	↓ 4.6	61.8	17.9	77.2	↑ 3.0
	DeepSeek-R1	95.3	↓ 0.7	57.0	↓ 0.2	64.6	↑ 0.7	84.1	↓ 6.8	50.6	13.8	66.9	↑ 3.4
	Qwen-3-235B	94.3	↓ 0.3	56.0	↓ 3.3	62.8	↓ 0.8	85.3	↓ 1.5	50.3	9.0	69.8	↑ 0.8

Table 1: Performance on the *Conflict* subset of NATCONFQA for non-reasoning (upper half) and reasoning models (lower half). Average precision, recall, and F_1 scores are reported for the two quality criteria (§4.3) — answer quality (left section) and conflict identification (right section). Results are shown for when models apply defaultive prompting (D), together with the absolute change in scores when applying conflict-aware prompting instead (ΔCA). Symbols ↓ / and ↑ / denote negative/significant-negative and positive/significant-positive changes. See Appendix H for details on significance-testing. Overall, conflict-aware prompting yields improvements for nearly all models across both evaluation criteria.

	Yes/No	WH-All	WH-Mix
defaultive	54.5 ± 2.2	53.1 ± 3.6	46.0 ± 5.7
conflict-aware	71.3 ± 2.0	72.9 ± 3.2	53.3 ± 5.6

Table 2: Average conflict-identification recall with 90% confidence intervals across eight models, reported for both prompting modes and all three conflict subsets (as detailed in §5.1).

5.2 Results on the *No-conflict* Subset

Next, we conducted an experiment which mirrors the traditional MAQA task, requiring models to generate responses that incorporate all relevant answers from the passages, *without* conflicting information. Specifically, we test four models' performance on the *No-conflict* subset of NATCONFQA, under both prompting modes, as a reference for the *Conflict* subset's experiments above. All tested models (both reasoning and non-reasoning LLMs) exhibit high answer quality under defaultive prompting ($F_{1\text{ans}} > 90$; see full results in Table 6 in Appendix F). This suggests that retrieving and integrating answers from passages is easier for models when no conflicts are present. However, when prompted in conflict-aware mode, performance slightly degrades (up to 5.5 in $F_{1\text{ans}}$), possibly because the enforced knowledge of potential conflicts (even when none exist) somewhat disrupts the model's natural inference.

5.3 Error Analysis – Dealing with Conflicts

The large performance gap between the *Conflict* and *No-conflict* subsets of NATCONFQA (§5.1, §5.2) calls for further examination. To that end, we conducted a manual error analysis on 160 sampled system responses generated by four models under defaultive prompting, on the *Conflict* subset (see Appendix I for full details). A human annotator categorized each system response into one of four main pre-defined error categories,[11] if an error was found (80 of the 160 instances), as follows:

Error	Description of response	Frequency
Choose	contains one reference answer	42%
Generalize	unifies answers by generalizing	17%
Resolve	hallucinates info to settle conflicts	13%
Refrain	does not answer question	5%

See examples in Table 9 in the appendix. The four models exhibit similar distributions of error types.

When the examined models made mistakes in their responses, it generally seems as if they tried to overcome conflicts through manipulative techniques. In 42% of the cases, they simply chose one answer in order to refrain from dealing with the conflicts. In 17% of the cases, they generated a response that unified the conflicting answers into a general answer that does not disclose the conflicts (e.g., by averaging numbers). In about 13%

---

[11] These categories were identified through a preliminary analysis for prominent error types, and inspired by Jiayang et al. (2024).

Dataset	Collection Method	#Instances	Conflicting Pairs	Conflict Type	Avg. #Passages	Avg. Passage Length (words)
ConflictingQA (Wan et al., 2024)	Automatic	238	All	Factual	9.2	314
WikiContradict (Hou et al., 2024)	Manual	253	All	Factual	2	43
DebateQA (Xu et al., 2024)	Automatic	2,941	N.D.	Point-of-View	4.2	4687.6
ECon (Jiayang et al., 2024)	Automatic	1,666	All	Factual	3	47.3
NATCONFQA (Ours)	Manual	677	All & Mixed & None	Factual	5.6	251.5

Table 3: Representative datasets for conflict-aware QA. "Conflicting Pairs" indicates whether an instance in the dataset has only conflicting answers (*All*), conflicting and non-conflicting answers (*Mixed*), no conflicting answers (*None*), or whether that distinction is not well defined *N.D.*. "Avg. # Passages" denotes the average number of passages per instance.

of the cases, the models fabricated information in an attempt to reconcile the conflicts. Finally, another approach was to simply respond with a general comment related to the question, without answering it.

These phenomena observed on high-end LLMs demonstrate the manners with which models attempt to overcome conflicting information. We call upon the research community to dive deeper into this matter, not only to solve conflict-related tasks such as ours, but also to better understand the way in which LLMs handle inconsistencies in knowledge. Future research should explore developing systems that embrace the complexity of conflicts rather than simply resolving them.

6 Related Work

In multi-answer QA, a question may have multiple valid answers, each supported by its own evidence (Voorhees, 2004). Although most datasets for this task generally assume that the different answers are consistent and complementary (Kwiatkowski et al., 2019; Zhu et al., 2020; Li et al., 2022; Zhong et al., 2022; Amouyal et al., 2023), in real-world scenarios, a query can expose conflicts or discrepancies between the different textual sources.

Yet, there are only several QA datasets that address conflicting answers, each exhibiting its own set of limitations. Table 3 shows the differences between NATCONFQA and existing benchmarks. In general, conflicting QA instances in naturally occurring texts are scarce, hence a popular strategy of prior works is to automatically introduce synthetic misinformation in texts, generating this way conflicting evidences for a QA instance (Jiayang et al., 2024; Liu et al., 2025b; Su et al., 2024; Wang et al., 2025; Ming et al., 2025). This approach inherently introduces artificial biases for the types of conflicts included in the dataset, as determined by the synthetic generation method. Another approach

to derive conflicting instances involves utilizing existing Yes/No questions coupled with documents retrieved from search results that contain conflicting information (Wan et al., 2024).

Other recent works focus on various types of conflicts in RAG settings, where there is a single correct answer (Liu et al., 2025a), multiple points of view (Xu et al., 2024), or a mix of different conflict types, such as temporal, misinformation, or opinion (Cattan et al., 2025). Most similar to our work, WikiContradict (Hou et al., 2024) includes human-annotated QAs that incorporate naturally-occurring (rather than synthetic) conflicting answers, found in Wikipedia articles. Yet, the instances in this dataset are limited to only two relatively short evidence passages, which always contradict each other.

In contrast to existing resources, NATCONFQA is a human-annotated dataset composed of naturally occurring conflicts between the different answers, covering both Yes/No and WH- questions. Additionally, each instance includes on average 5.5 passages. Importantly, our work is the first to collect fine-grained annotations for each pair of answers, where some answer pairs are conflicting while others are not (the "mixed" category in Table 3. This annotation scheme enables more realistic assessment of models' ability to identify naturally occurring conflicting answers, while distinguishing them from non-conflicting answers.

7 Conclusion

In this work we enhance the *Conflict-Aware Multi-Answer QA* task by explicitly requiring conflict identification among answers. We create a dataset for the enhanced task, via a novel cost-effective methodology that leverages fact-checking datasets. Our NATCONFQA dataset is a realistic, conflict-rich benchmark that challenges current strong models. We test several state-of-the-art LLMs on the dataset, and show that models still struggle with surfac-

ing conflicting answers consistently, particularly in instances that contain both conflicting and non-conflicting answers, even when expressly prompted to be on the watch for potential conflicts. Finally, an error analysis of model responses exposes manners in which LLMs mishandle conflicts.

Limitations

We employed LLMs for many tasks throughout this paper, including conflict-aware MAQA, evaluation of several criteria, and response decomposition. While we conducted some reasonable prompt-engineering for these assignments, it is possible that even more effective prompts would improve or change results.

Since pre-trained LLMs' training datasets are not fully documented, we can't rule out overlap with the underlying data used for creating our dataset, raising the risk of contamination.

References

Rami Aly, Christos Christodoulopoulos, Oana Cocarascu, Zhijiang Guo, Arpit Mittal, Michael Schlichtkrull, James Thorne, and Andreas Vlachos, editors. 2021. *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*. Association for Computational Linguistics, Dominican Republic.

Samuel Amouyal, Tomer Wolfson, Ohad Rubin, Ori Yoran, Jonathan Herzig, and Jonathan Berant. 2023. QAMPARI: A benchmark for open-domain questions with many answers. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 97–110, Singapore. Association for Computational Linguistics.

Arie Cattan, Alon Jacovi, Ori Ram, Jonathan Herzig, Roee Aharoni, Sasha Goldshtein, Eran Ofek, Idan Szpektor, and Avi Caciularu. 2025. Dragged into conflicts: Detecting and addressing conflicting sources in search-augmented llms.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *CoRR*, abs/2412.19437.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2021. Climate-fever: A dataset for verification of real-world climate claims.

Kevin Fischer, Darren Fürst, Sebastian Steindl, Jakob Lindner, and Ulrich Schäfer. 2024. Question: How do large language models perform on the question answering tasks? answer:.

Yufang Hou, Alessandra Pascale, Javier Carnerero-Cano, Tigran Tchrakian, Radu Marinescu, Elizabeth Daly, Inkit Padhi, and Prasanna Sattigeri. 2024. Wikicontradict: A benchmark for evaluating llms on real-world knowledge conflicts from wikipedia.

Cheng Jiayang, Chunkit Chan, Qianqian Zhuang, Lin Qiu, Tianhang Zhang, Tengxiao Liu, Yangqiu Song, Yue Zhang, Pengfei Liu, and Zheng Zhang. 2024. Econ: On the detection and resolution of evidence conflicts.

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

Haonan Li, Martin Tomko, Maria Vasardani, and Timothy Baldwin. 2022. MultiSpanQA: A dataset for multi-span question answering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1250–1260, Seattle, United States. Association for Computational Linguistics.

Siyi Liu, Qiang Ning, Kishaloy Halder, Zheng Qi, Wei Xiao, Phu Mon Htut, Yi Zhang, Neha Anna John, Bonan Min, Yassine Benajiba, and Dan Roth. 2025a. Open domain question answering with conflicting contexts. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1838–1854, Albuquerque, New Mexico. Association for Computational Linguistics.

Siyi Liu, Qiang Ning, Kishaloy Halder, Wei Xiao, Zheng Qi, Phu Mon Htut, Yi Zhang, Neha Anna John, Bonan Min, Yassine Benajiba, and Dan Roth. 2025b. Open domain question answering with conflicting contexts.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

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

Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. 2025. Faitheval: Can your language model stay faithful to context, even if "the moon is made of marshmallows". In *The Thirteenth International Conference on Learning Representations*.

OpenAI-Team. 2024a. Gpt-4 technical report.

OpenAI-Team. 2024b. Gpt-4o system card. *CoRR*, abs/2410.21276.

John W. Pratt. 1959. Remarks on zeros and ties in the wilcoxon signed rank procedures. *Journal of the American Statistical Association*, 54(287):655–667.

Mourad Sarrouti, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. 2021. Evidence-based fact-checking of health-related claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3499–3512, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. 2024. Conflictbank: A benchmark for evaluating the influence of knowledge conflicts in llm.

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

Ellen M. Voorhees. 2004. Overview of the trec 2003 question answering track. In *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003)*, Gaithersburg, MD. National Institute of Standards and Technology.

Alexander Wan, Eric Wallace, and Dan Klein. 2024. What evidence do language models find convincing?

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2025. Retrieval-augmented generation with conflicting evidence.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

Rongwu Xu, Xuan Qi, Zehan Qi, Wei Xu, and Zhijiang Guo. 2024. Debateqa: Evaluating question answering on debatable knowledge.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *CoRR*, abs/2306.05685.

Victor Zhong, Weijia Shi, Wen tau Yih, and Luke Zettlemoyer. 2022. Romqa: A benchmark for robust, multi-evidence, multi-answer question answering.

Ming Zhu, Aman Ahuja, Da-Cheng Juan, Wei Wei, and Chandan K. Reddy. 2020. Question answering with long multiple-span answers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3840–3849, Online. Association for Computational Linguistics.

A Fact-Checking Datasets for Conflict-aware MAQA

We build upon two established fact-checking benchmarks, CLIMATE-FEVER and HEALTHVER, which provide real-world claims paired with supporting, refuting, and neutral evidence passages.

CLIMATE-FEVER. The CLIMATE-FEVER dataset (Diggelmann et al., 2021) includes 808 claims. Of these, 654 claims have only non-conflicting evidence (all support), while 154 claims are labeled as *disputed*, containing conflicting evidences.

HEALTHVER. The HEALTHVER dataset (Sarrouti et al., 2021) comprises 1,084 claims. Among them, 607 claims feature only supporting evidence (no conflict), and 477 claims include both supporting and refuting evidence, yielding true conflicts.

To ensure a diverse and representative subset of questions, we address the high redundancy in HEALTHVER, where many real-world health-related claims are duplicated. We randomly sampled a single instance per topic from the 477 conflicting claims spanning on 55 topic questions. This process resulted in a final subset of 55 unique instances.

Licensing. Neither CLIMATE-FEVER nor HEALTHVER specify an explicit license. Upon publication, we will release NATCONFQA under the CC BY 4.0 license,[12] permitting unrestricted reuse with attribution for research purposes.

B Dataset Annotation Details

Annotator details. Our two annotators were undergraduate or graduate students, and are fluent English speakers. They underwent two training

---

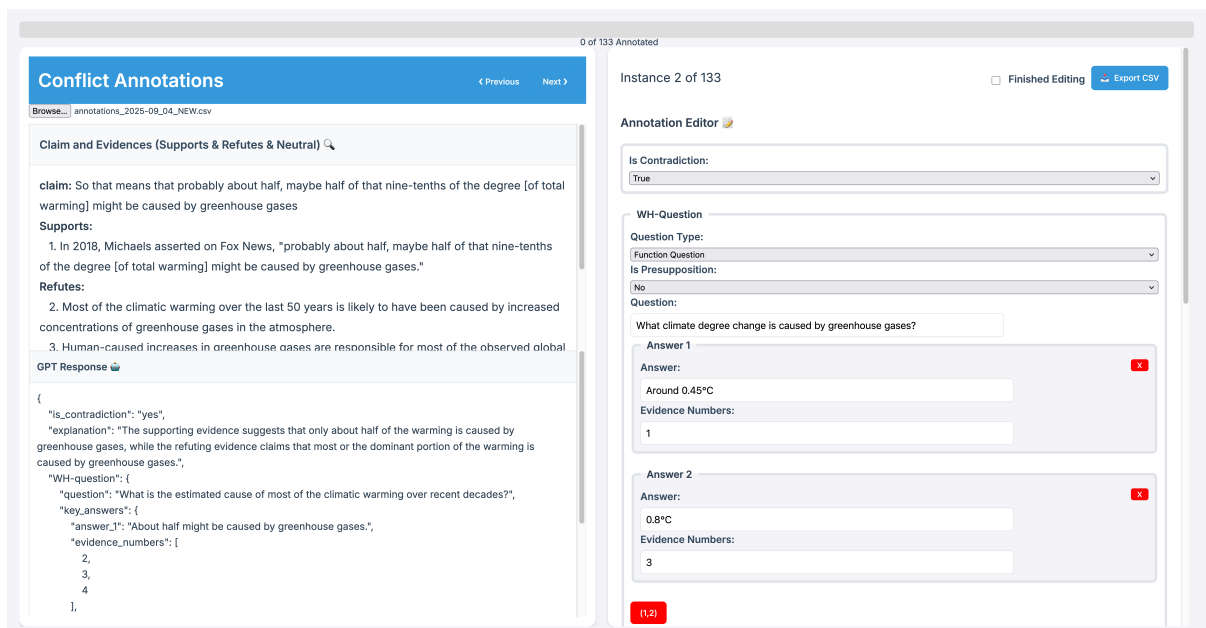[12] https://creativecommons.org/licenses/by/4.0/

Figure 2: Screenshot of our annotation interface. On the left, annotators view the fact-checking instance, including the claim and evidence sentences grouped by their initial labels (support, refute, neutral). At the bottom, LLM-generated WH and yes/no question–answer suggestions are displayed (see Figure 9 for the prompts). On the right, annotators write or edit their own questions and corresponding answers.

iterations, each on 10 instances. They were compensated at approximately $14 per hour. The annotators worked a combined total of 52 hours to prepare the NatConfQA dataset. The annotators were informed that their annotations are for research purposes, and that they can terminate their participation in the process whenever they want.

Annotation tool. To facilitate the annotation process, we developed a dedicated annotation tool that supports question-writing, linking respective evidence, and labeling conflicting answer pairs.

We present screenshots of the annotation interface and guidelines for both WH and Yes/No questions in Figures 2 and 3, respectively. The provided instructions guided annotators in accurately identifying conflicts and extracting relevant answers based on the given evidence. Additionally, annotators were instructed to record the evidence IDs that support each annotated answer. Our custom annotation tool further allows annotators to add as many answers as necessary and seamlessly author both WH-type and binary (Yes/No) questions for each claim. While the guidelines place a strong emphasis on handling conflict instances, annotators were allowed to skip the conflict-oriented instructions when working on the support instances.

C Passage Extraction for Evidence Sentences

To simulate a realistic QA setting, we reverse the typical sentence-level focus of fact-checking datasets by retrieving entire passages surrounding each evidence span.

For HealthVer, which is derived from scientific abstracts in the CORD-19 corpus[13], we locate the original abstract corresponding to each evidence sentence. We perform an exact string match of the sentence within the CORD-19 collection and extract the full abstract to serve as the passage context.

For climate-fever, which uses evidence drawn from Wikipedia, we scraped the English Wikipedia pages as of February 1, 2020 and converted them to plain text. We then employ the RapidFuzz fuzzy-matching library[14] to identify each evidence sentence within its article. Finally, we heuristically expand to the surrounding paragraph—defined by nearest blank lines or section headers—to create a coherent passage that preserves the original narrative flow.

This passage-level extraction ensures that each QA instance reflects the broader context in which evidence appears, aligning our setup with realistic

---

[13] https://www.semanticscholar.org/cord19
[14] https://github.com/maxbachmann/RapidFuzz

Statistic	*No-conflict*	*Conflict*
# Instances	408	269
Avg. # Passages	6.7	5.6
Avg. # Answers (WH) [min–max]	3.8 [2–13]	2.6 [2–10]
Avg. # Conflict Pairs	–	1.3

Table 4: Statistics for our NATCONFQA dataset. The three average stats are per instance. *Avg. # Passages* is the mean number of passages per instance; *Avg. # Answers (WH) [min–max]* reports the average number of answers per WH question and the corresponding range; *Avg. # Conflict Pairs* is the average number of conflicting answer pairs per instance (e.g., always 1 for Yes/No).

Model Name	Model Tag	Reasoning
GPT-4o	gpt-4o-2024-08-06	✗
Gemini 2.0 Flash	gemini-2.0-flash-001	✗
DeepSeek-V3	deepseek-ai/DeepSeek-V3	✗
Qwen-2.5-72B	Qwen/Qwen2.5-72B-Instruct-Turbo	✗
o3	o3-2025-04-16	✓
Gemini 2.5 Pro	gemini-2.5-pro-preview-03-25	✓
DeepSeek-R1	deepseek-ai/DeepSeek-R1	✓
Qwen-3-235B	Qwen/Qwen3-235B-A22B-fp8	✓
o4-mini	o4-mini-2025-04-16	✓

Table 5: Exact model version tags for the models used in this work.

Retrieval-Augmented Generation workflows.

D NATCONFQAStatistics

Table 4 presents statistics for NATCONFQA,[15] including number of questions, and number of passages, answers and conflicts per question. The dataset comprises two subsets, one with only supporting answers (titled *No-conflict* in the table) and another with conflicting answers (titled *Conflict*). For both types, there are WH-questions as well as Yes/No questions.

E Details for the Experimental Setup

In this section, we provide technical details of the experimental setup defined in Section 4.2. All experiments were conducted between May 1 and May 19, 2025, using the OpenAI,[16] Google,[17] and Together.ai[18] APIs. The exact version tags for all models utilized in this work are listed in Table 5.

For reproducibility, we set the temperature to 0 for all models with a maximum number of generated tokens (max_tokens) at 512. For the

---

Model	Answer Quality					
	$\text{precision}_{\text{ans}}$		$\text{recall}_{\text{ans}}$		$F_{1\text{ans}}$	
	D	ΔCA	D	ΔCA	D	ΔCA
non-reasoning						
GPT-4o	98.6	↓ 1.4	95.6	↑ 0.7	95.9	↓ 0.1
DeepSeek-V3	96.9	↓ 1.1	95.5	↑ 0.1	95.6	↓ 1.0
reasoning						
o3	98.3	↓ 1.5	97.2	0.0	97.5	1.2
DeepSeek-R1	97.6	↓ 1.3	89.7	↓ 4.2	90.7	5.5

Table 6: Average precision, recall, and F_1 percentages for the answer-quality criterion (§4.3) on the *No-conflict* subset of NATCONFQA, reported for four LLMs. All models exhibit high scores (> 89) on all measures. See the caption of Table 1 for further details on notations.

OpenAI models (o3 and o4-mini), we configure reasoning_effort to high. For the Gemini models, we set thinking_budget to 1024 for Gemini 2.5 Pro and to 0 for Gemini 2.0 Flash. The total cost of the experiments using the three LLM APIs was approximately $500.

The prompts used for model evaluation are shown in Figure 4, while those for the LLM-as-a-judge are presented in Figure 5. There are five prompts in total for the LLM-as-a-judge: one for answer decomposition (Prompt 1) and four for the binary decision functions defined in Section 4.3 (Prompts 2–5), namely, $\mathcal{J}_{\text{ans}}(a_i, y)$, $\mathcal{J}_{\text{ans}}(\hat{a}_i, P)$, $\mathcal{J}_{\text{conf}}(a_i, a_j, y)$, and $\mathcal{J}_{\text{conf}}(\hat{a}_i, \hat{a}_j, P)$.

F NATCONFQA *No-conflict* Subset Results

In Table 6, we present results on the *No-conflict* subset of NATCONFQA, using the same metrics and notation conventions as in Table 1.

G Effectiveness of LLM-as-a-Judge

Manually annotating answer quality and conflict identification is expensive, so we rely on a fast automatic judge — o4-mini[19] — and validate its agreement with humans. we randomly sampled 120 system responses from two models (o3-high and Gemini 2.5 Pro) under both prompting modes (§4.1). A human judge then applied the judgment protocol explained in Section 4.3, with guidelines similar to the instructions in the prompts in Figure 5, decomposing answers and making the binary decisions that constitute both metrics.

To assess the reliability of the automatic judge, we compute Pearson correlations between its la-

---

[15]For randomly sampled instances from the dataset, see Table 10.

[16]https://platform.openai.com

[17]https://aistudio.google.com

[18]https://together.ai

[19]https://openai.com/index/o3-o4-mini-system-card

Evaluation Metric	Recall	Precision
Answer quality	0.6645	0.6221
Conflict identification	0.6744	0.6371

Table 7: Pearson's r correlation coefficients for recall and precision, computed between human and LLM-as-judge over 120 samples ($p < 0.00003$). The results show strong correlations for all metrics between human and LLM judgments.

bels and the human annotations (Table 7). Both metrics show a strong positive correlation, with all correlation values exceeding 0.6.

Furthermore, the human evaluation results (Table 8) show performance trends similar to those from the automatic evaluation (Table 1). In both cases, conflict-aware prompting leads to higher recall — particularly for conflict identification.

H Significance Testing on Results

To compute significance in Table 1, we conducted Wilcoxon signed-rank tests comparing the default-ive and conflict-aware modes (Wilcoxon, 1945). We applied Pratt's conservative zero-difference method (Pratt, 1959) and report significance at $p < 0.01$.

I Error Analysis Details

In order to assess model errors, we sampled 160 outputs, using the default prompt described in §5.1, from four models (O3, DeepSeek-R1, GPT-4o, and DeepSeek-V3). Annotation was performed by a graduate student at an hourly rate of $14. We sampled 40 instances for each model where 20 were Yes/No questions and 20 were WH-question.

J Conflict Subsets Analysis

Figure 7 compares conflict-identification recall across three subsets—Yes/No, WH-All (all answer pairs conflict), and WH-Mix (mixture of conflicting and non-conflicting pairs)-for each model under two prompting types: Defaultive and Contradict-Aware. Bars show mean recall per model. Two consistent trends emerge. First, Contradict-Aware prompting substantially improves performance for most models, especially on Yes/No and WH-All, while gains on WH-Mix are smaller. Second, WH-Mix is the hardest subset: it has the lowest base recall under Defaultive prompting and remains lowest even after Contradict-Aware prompting. This suggests that instances containing both conflicting and non-conflicting evidence introduce challenge that current models struggle to resolve.

K AI Assistance

Throughout this project, we were assisted by AI tools to accelerate both code implementation (some code snippets) and manuscript preparation (local rephrasing). We carefully reviewed and refined all AI-generated content to ensure technical accuracy and stylistic consistency.

Model	Answer Quality						Conflict Identification					
	$\text{precision}_{\text{ans}}$		$\text{recall}_{\text{ans}}$		$F_{1\text{ans}}$		$\text{precision}_{\text{conf}}$		$\text{recall}_{\text{conf}}$		$F_{1\text{conf}}$	
	D	ΔCA	D	ΔCA	D	ΔCA	D	ΔCA	D	ΔCA	D	ΔCA
o3	85.5	↓ 0.8	84.8	5.3	85.1	↑ 2.7	100.0	↓ 6.7	46.7	26.7	85.7	↑ 4.3
gemini-2.5-pro	79.9	↓ 0.6	83.4	↑ 5.3	81.6	↑ 1.9	81.8	↑ 11.5	46.7	33.3	57.1	↑ 42.9

Table 8: Human evaluation of performance on the *Conflict* subset of NATCONFQA for the two top performing models from Table 1, based on 120 instances. Average precision, recall, and F_1 scores are reported (percent) for the two evaluation criteria (§4.3): answer quality (left) and conflict identification (right). Columns show results for default prompting (D) and the absolute change when using conflict-aware prompting (ΔCA). Symbols ↓/ and ↑/ denote negative/significant negative and positive/significant positive changes, respectively (using $p < 0.05$, see Appendix H for significance testing methodology).

Error Type	Description	Example (Question / Reference answers / Model answer)
Choose Answer	The model outputs only one of the answers, omitting the others.	**Q:** What is the rate of ice mass loss in Antarctica? **Reference:** "+82 Gt/yr" and "–220 Gt/yr" **Model:** "Antarctica is gaining 82 Gt of ice per year."
Answer Generalization	The model summarizes conflicting values vaguely (e.g., as an average or range) instead of presenting them distinctly.	**Q:** How much has global temperature risen? **Reference:** "0.45 °C" and "0.8 °C" **Model:** "Temperatures have increased between 0.45 to 0.8."
Conflict Resolution	The model mentions all answers but presents them as if there is no conflict, possibly with hallucinatory information.	**Q:** What is the estimated global temperature rise since 1900? **Reference:** "0.45 °C" and "0.8 °C" **Model:** "Until 1945, the rise was 0.45 °C, and then 0.8 °C."
Refrain from answering	The model does not provide an answer to the question or provides an irrelevant response.	**Q:** What is the estimated rate of ice loss from Greenland per year? **Reference:** "Between 200 and 300 Gt/yr" and "Approximately 220 Gt/yr" **Model:** "Greenland is a large landmass covered in ice."

Table 9: Error types for the error analysis. Each instance is labeled with a single error type based on the model's ability to reflect, miss, or misrepresent the underlying conflict.

NATCONFQA Manual Annotation Guidelines

1. Task Overview

- Our goal is to reveal conflicts in data through questions and answers.

- You will be provided with a claim along with supporting, refuting, and neutral evidence.

- Your task involves two main objectives:

  (1) **Conflict detection**: Assess whether there is a conflict between the supporting and refuting evidence.
      Yes   No   Uncertain

  (2) **Q–A Generation**: Formulate questions that capture the conflict based on the given claim and evidence. For each question, provide the differing answers along with the evidence IDs supporting each answer.

2. Rules for Effective Question Formulation

You should follow the following rules when writing the questions:

1. **Conflict-inducing** – The question should prompt a conflicting response.

2. **Specificity** – Ensure the question targets detailed aspects of the text rather than general topics.

3. **Assumption-free** – The question should be free from specific assumptions (from the text).

3. Rules for Effective Answer Formulation

You should follow the following rules when writing the answers:

1. **Completeness** – Ensure that the answers address all the evidence provided.

2. **Conciseness** – Keep the answers brief and to the point while maintaining clarity.

3. **Relevance** – Ensure that the answers directly address the question asked.

4. **Atomicity** – Each answer should contain only a single response. If multiple answers exist, they should be separated into distinct answers rather than combined.

Figure 3: Summary of the annotation guidelines for creating our NATCONFQA dataset. The full guidelines will be provided along with the complete resources.

Source Data	Q-Type	A-Type	Question	Answers
CLIMATE-FEVER	YN	No-conflict	Do changes in land use, such as agriculture and deforestation, contribute to Earth's climate changes?	1. Yes
CLIMATE-FEVER	YN	Conflict	Do greenhouse gases increase Earth's temperature?	1. Yes 2. No
CLIMATE-FEVER	wh	No-conflict	What are the ranges of water vapor's contribution to the greenhouse effect?	1. Between 36% and 66% under clear sky conditions 2. Between 66% and 85% when including clouds
CLIMATE-FEVER	wh	Conflict	What is the perceived level of scientific agreement on climate change?	1. There is a scientific consensus on human-caused climate change 2. Scientific opinion is evenly divided or completely unsettled
HEALTHVER	YN	No-conflict	Does favipiravir decrease viral replication in COVID-19 patients?	1. Yes
HEALTHVER	YN	Conflict	Do natural remedies help prevent you from getting infected with COVID-19?	1. Yes 2. No
HEALTHVER	wh	No-conflict	What are the clinical symptoms observed in patients with COVID-19?	1. Fever 2. Dry cough 3. Sore throat 4. Dyspnea 5. Fatigue 6. Myalgia 7. Headache 8. Loss of smell and taste
HEALTHVER	wh	Conflict	What is possible in regards to cat to human coronavirus transmission?	1. Cats can transmit the virus to humans 2. Cats cannot transmit the virus to humans

Table 10: Eight randomly sampled instances from NATCONFQA. Q-Type: questoin type YN = yes/no; wh = WH-questions. A-Type: answer type. When the dataset is published, it will also include the associated passages supporting each answer. For simplicity, and since conflict instances in this sample contain only two answers, we do not include which specific answer pairs are in conflict.

NATCONFQA Dataset Prompt Template

Provide a concise, single-sentence answer that includes every distinct answer to the following question, based on the given passages from multiple sources. If any answers conflict, clearly indicate which ones are in conflict while remaining objective and neutral.

Question: {question}
Passages: Passage #1:
{Passage 1 text}

Passage #2:
{Passage 2 text}
. . .

Figure 4: The prompt template used to prepare inputs for the LLMs when testing on NATCONFQA. This is the template for both modes (defaultive prompting mode and conflict-aware prompting mode) as defined in §4.1. The underlined sentence is omitted in the defaultive mode and included in the conflict-aware mode.

LLM-as-a-Judge Prompt Templates

1. System Response Decomposition into Distinct Answers $\hat{A}$

Given the following question and long-form system answer, extract all distinct short answers that directly address the question. Ensure to extract all such sub-answers. The output must be a JSON object with a `"sub_answers"` key whose value is an array of sub-answer strings.

Question: "{question}"
System answer: "{model_answer}"

2. Answer a_i Exist in System Response y

Determine if the reference short answer appears anywhere in the longer system answer — even if it's noted as incorrect, anecdotal or implicitly mentioned. Return a JSON object with a single key, `"answer"`, whose value is either `"true"` or `"false"`:

- `"true"` if the reference short answer is present in any form (explicit or implicit).

- `"false"` otherwise.

Question: "{question}"
Reference short answer: "{reference_answer}"
System answer: "{model_answer}"

3. Answer $\hat{a}_i$ Exists in the passages P

Determine if the given short answer appears anywhere in the following passages — even if it's noted as anecdotal or implicitly mentioned. Return a JSON object with a single key, `"answer"`, whose value is either `"true"` or `"false"`:

- `"true"` if the given short answer is present in any form (explicit or implicit).

- `"false"` otherwise.

Question: "{question}"
Short answer: "{model_answer}"
Passages: {passages}

4. Answers a_i and a_j Conflict Within System Response y

Your task is to assess whether a pair of short answers to a question are clearly indicated as conflicting answers in a longer answer written by a question answering system. You should return a JSON object with an `"answer"` key `"true"` if the pair of short answers clearly indicated as conflicting in the longer system's answer or `"false"` otherwise.

Question: "{question}"
Short answer #1: "{short_answer1}"
Short answer #2: "{short_answer2}"
System answer: "{model_answer}"

5. Answers $\hat{a}_i$ and $\hat{a}_j$ Conflict Within Passages P

Your task is to assess whether the two short answers, which are located in the following passages, are clearly indicated as conflicting to the same question. You should return a JSON object with an `"answer"` key `"true"` if the pair of short answers clearly indicated as conflicting or `"false"` otherwise.

Question: "{question}"
Short answer #1: "{short_answer1}"
Short answer #2: "{short_answer2}"
Passages: {passages}

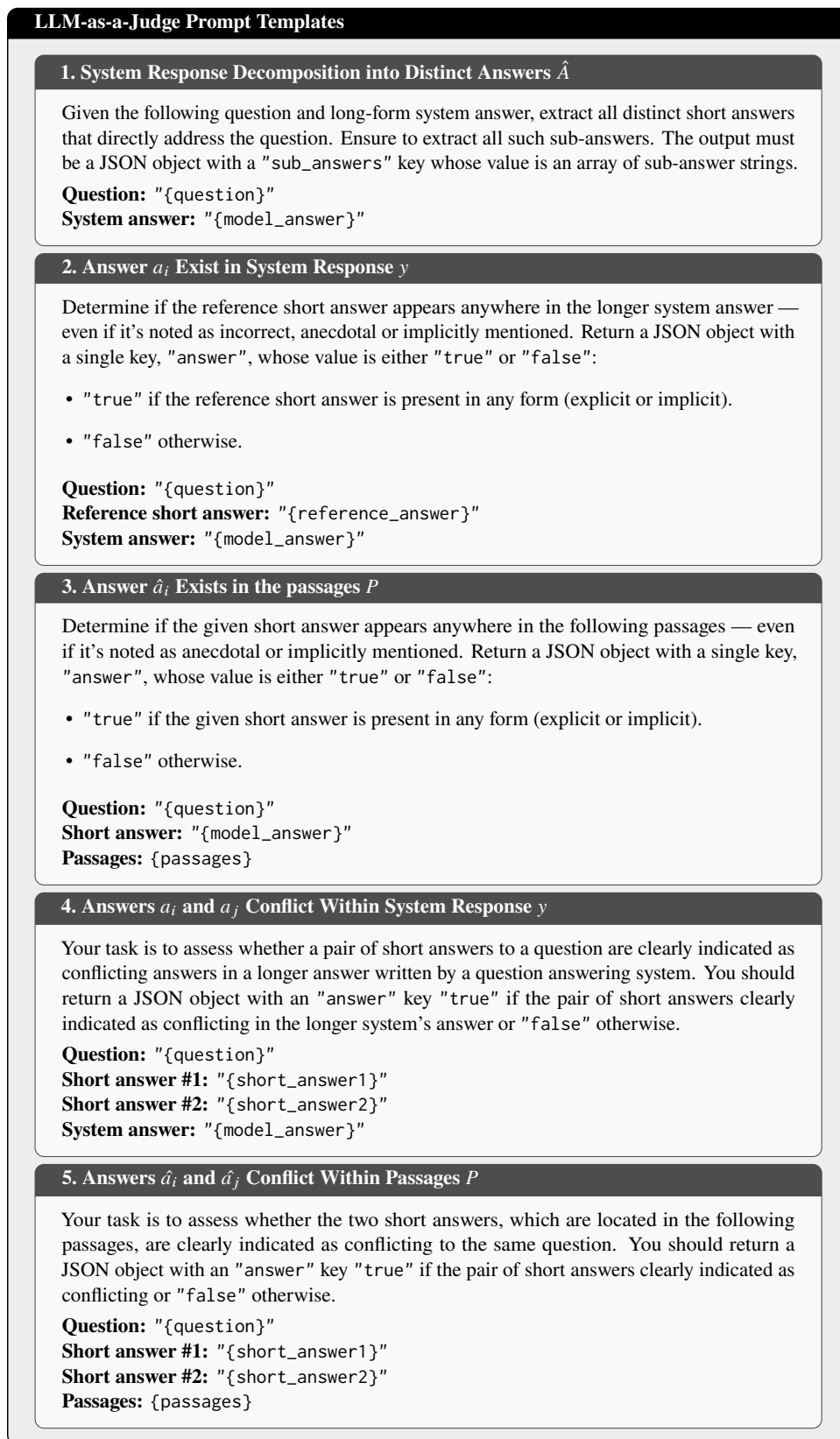Figure 5: Prompt templates used in our experiments for LLM-as-a-judge as detailed in Section 4.3.

Manual Annotation Guidelines for Correlation Assessment

Task Overview: You are given a question, human-annotated reference answers (referred to as *reference_answers*), human-annotated conflicting answer pairs (referred to as *reference_conflicting_answer_pairs*), associated passages, and a model-generated answer; please follow the steps below.

1. Decompose the Model Answer into Atomic Answers

- Read the `model_answer` and split it into independent facts or claims ("atomic answers").

- Number them 0, 1, 2, . . . in the order they appear.

- Enter them in the format:

  - 0: First atomic answer. . .
  - 1: Second atomic answer. . .
  - 2: Third atomic answer. . .

- → Fill in `Model Answers Decomposed`.

2. Match Decomposed Answers to Reference Answers

- For each atomic answer (by index), check if it appears in the `reference_answers`.

- List the indices of those that match, e.g., `[0, 2]`.

- → Fill in `Matched Answers in Reference Answers`.

3. Identify Conflicting Reference-Answer Pairs

- Review the `reference_conflicting_answer_pairs` and identify each pair that directly contradicts the other.

- In the model's response, look for contrastive cues like "however," "but," or "on the other hand."

- Record each conflicting pair by their indices, e.g., `[(0,1), (1,2)]`.

- → Fill in `Conflicting Reference Answer Pairs Found`.

4. Match Model Answers to Passages or References

- For each atomic answer, check if it is supported by either a `reference_answer` or a paragraph.

- List the indices of atomic answers that are supported, e.g., `[0, 1]`.

- → Fill in `Found Model Answers in Reference Answers/Passages`.

5. Identify Conflicts among Model's Answers

- For each pair of atomic answers flagged by the model as conflicting, record:

  - `(i, j): 1` if the conflict is correct.
  - `(i, j): 0` if the conflict is incorrect.

- Example: `(0, 2): 1, (0, 1): 0`

- → Fill in `Found Conflicting Model's Answers`.

Figure 6: Annotation guidelines used by human annotators to assess the correlation between automatic LLM-based judgments and human evaluation. The process includes both conflict identification and answer quality metrics, supporting the measurement of recall and precision for each.
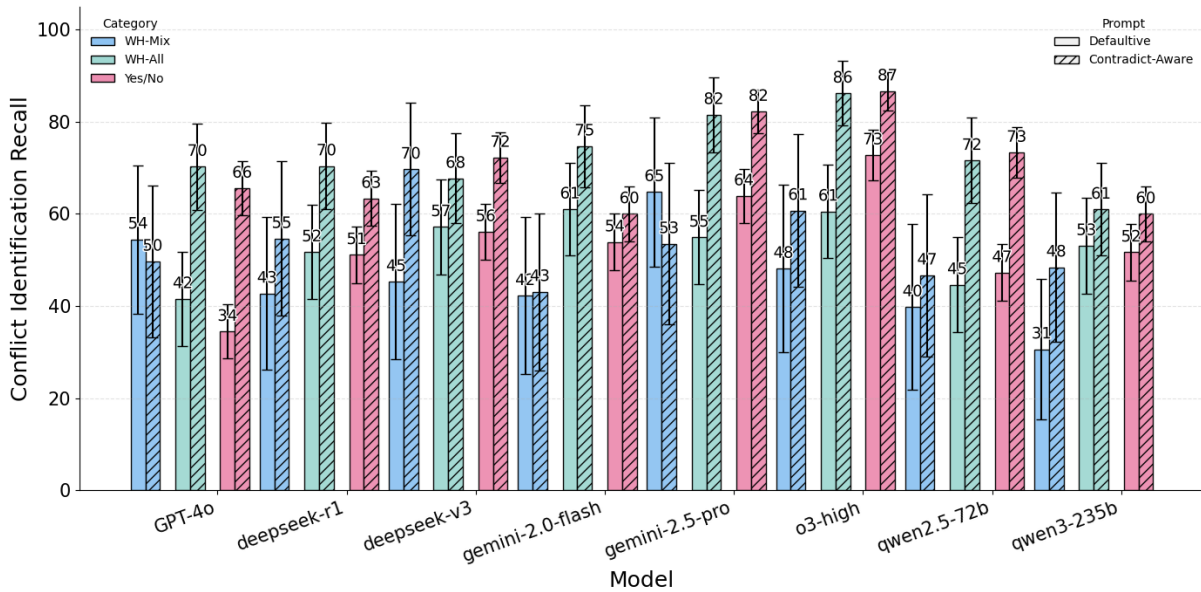
Figure 7: Conflict Identification Recall across models on three instance subsets (WH-Mix, WH-All, Yes/No), split by prompt type (defaultive vs. contradict-aware). Bars show mean conflict identification recall with 90% confidence intervals. Overall, contradict-aware consistently improves recall relative to defaultive (Table 1); however, performance on WH-Mix is notably lower than on WH-All and Yes/No for nearly all models, and the gains from contradict-aware are smaller on this subset, highlighting the challenge when conflicting and non-conflicting signals co-occur within the same instance.

Manual Annotation Guidelines for Error Analysis

Task Overview: You are provided with a question, the model's answer for that question, and a set of annotated reference answers. First, review the list of error categories in Table 9. Then, for each model output, classify it into the single most appropriate category, based on the explanation that best matches the model's output.

Steps:

- Read the **question**, **reference answers**, and the **model's answer**.

- If the model presents both sides with clear contrast (e.g., "but," "however"), label it -2.

- Otherwise, choose a single error ID that best describes the model error.

- Use -1 if the error does not fit any predefined category and describe it in the notes.

Note: Annotate each instance with only one ID. Include a brief justification if necessary.

Figure 8: Annotation guidelines used by human annotators for conducting error analysis on outputs from different models.

Prompt for NATCONFQA Question and Answers Suggestion for Conflicting Instances

You are given a claim along with multiple pieces of evidence, categorized as supporting, refuting, or neutral in relation to the claim. Your task is to analyze the evidence and perform the following subtasks:
(1) **Determine Contradiction:** Identify whether there is a direct contradiction between the supporting and refuting evidence. Respond with ″yes″ if such a contradiction exists, or ″no″ if not.
(2) **Explain Contradiction (if applicable):** Briefly explain how the supporting and refuting evidence conflict.
(3) **Generate a WH-Question (if applicable):** Create a factoid-style WH-question based on the conflicting information. Avoid directly referencing the contradiction. Provide short answers and indicate the evidence numbers supporting each.
(4) **Generate a Yes/No Question:** Frame a yes/no question that reflects the core contradiction or claim, supported by evidence.
Output Format:

```
{
  "is_contradiction": "yes" or "no",
  "explanation": "...",
  "WH-question": {
    "question": "...",
    "key_answers": {
      "answer_1": "...",
      "evidence_numbers": [...],
      ...
    }
  },
  "Yes/no-question": {
    "question": "...",
    "Yes_answers": [...],
    "No_answers": [...]
  }
}
{instance}
```

Prompt for NATCONFQA Question and Answers Suggestion for Non-conflicting Instances

You are given a claim along with multiple pieces of evidence, categorized as supporting or neutral. Your task is to:
(1) **Generate a WH-Question:** Create a concise WH-question based on the supporting evidence. Provide short answers and list the supporting evidence numbers.
(2) **Generate a Yes/No Question:** Frame a yes/no question grounded in the supporting evidence and reflecting the claim's context.
Output Format:

```
{
  "WH-question": {
    "question": "...",
    "key_answers": {
      "answer_1": "...",
      "evidence_numbers": [...],
      ...
    }
  },
  "Yes/no-question": {
    "question": "...",
    "Yes_answers": [...],
    "No_answers": [...]
  }
}
{instance}
```

Figure 9: Prompt templates used to generate suggested questions and answers for annotators in the NATCONFQA annotation process. The first prompt is applied to instances containing conflicting evidence, while the second is used for non-conflicting evidence.

Annotation Guidelines for NᴀᴛCᴏɴꜰQA Dataset Quality Assessment

1. Task Overview

- You will be provided with:

    - A **question** prompting the model's response
    - A list of **answers**, each associated with a unique ID
    - A list of **evidence** that may support the answers
    - A list of **answer pair labeled as conflicting**, specified by their answer IDs (if any)

- Your task consists of two main steps:

  (1) **Answer Quality:** For each answer, determine whether it is clearly supported by at least one of the provided evidence.
    - Mark 1 if it is grounded (i.e., the answer is directly supported by any evidence).
    - Mark 0 if none of the evidence supports the answer.
  (2) **Conflict Detection:** For each answer pair labeled as conflicting, determine whether they conflicting each other based on the provided evidence.
    - Mark 1 if the two answers clearly conflicting each other.
    - Mark 0 if the answers are compatible or describe different aspects that can co-exist.

2. Notes and Clarifications

- Focus only on what is explicitly stated in the evidence and answers. Avoid making assumptions beyond the given text.

- If a pair appears borderline or ambiguous, lean toward 0 and leave a short note.

- Use the comments field to explain any unclear cases or edge scenarios you encounter.

Figure 10: Annotation guideline to evaluate the quality of NᴀᴛCᴏɴꜰQA. We assess both answer quality and conflict identification using an external annotator. The annotation procedure includes verifying whether answers are supported by evidence and whether identified answer pairs are truly in conflict. Results show high agreement, indicating the task is well-defined and the annotation protocol is reliable.