Bridging the Logic Gap in Fact-Checking through Content–Logic Coupled Evaluation Paradigm

Anonymous ACL submission

Abstract

Driven by the rise of social media and generative artificial intelligence, the internet has been flooded with unverified information, making reliable fact-checking increasingly critical. Most existing fact-checking research adheres to the decompose-then-verify paradigm, emphasizing verification of individual facts while overlooking the validity of logical dependencies among them. As a result, text containing logical errors may still be misjudged as factual. Moreover, existing datasets and metrics focus on fact completeness and coverage, failing to capture the logical dimension. To help bridge this gap, we propose a content–logic coupled factuality evaluation paradigm, which conceptualizes factuality along two complementary dimensions: content factuality and logic factuality. Under this paradigm, we introduce a holistic solution consisting of LoReFact, the first long-form fact-checking dataset that systematically incorporates the logical dimension; LoRe-Factcheck, a simple yet effective framework for joint content–logic evaluation; and a logic-aware metric named LoReFactScore for exposing and penalizing logical fallacies. Experiments demonstrate the importance of logical factuality and the effectiveness of our proposed paradigm for fact-checking.

1 Introduction

With the widespread deployment of large language models (LLMs) and the growth of online information platforms, the internet has become increasingly populated with content generated by both human users and LLMs. To attract attention or enhance persuasiveness, some of this content blends partially correct information with misleading narrative logic, posing new challenges for fact-checking (Huang et al., 2025; Ji et al., 2023; Zhang et al., 2023).

Most existing fact-checking research adopts the decompose-then-verify paradigm. The core idea is to decompose the text into atomic claims, verify them individually against external knowledge sources, and aggregate the resulting claim-level judgments into an overall factuality score for the text. Following this paradigm, prior studies mainly focus on optimizing individual steps of the pipeline, such as improving the extraction of atomic claims (Min et al., 2023; Song et al., 2024; Gunjal and Durrett, 2024; Fatahi Bayat et al., 2025), and employing LLMs as agents for retrieval and verification (Chern et al., 2023; Wei et al., 2024).

However, the decompose-then-verify paradigm mainly focuses on individual atomic claims and overlooks the logical dependencies among them. Intuitively, this leads to a potential issue: when partially correct facts are combined through incorrect logical relationships, existing methods may misjudge the text containing logical fallacies as factual, as illustrated in Figure 1 (a). To quantify this issue, we conducted a small-scale preliminary study using 50 logic-related samples collected from PolitiFact[1]. We evaluated three representative methods, SAFE (Wei et al., 2024), VeriScore (Song et al., 2024), and FactCheck-GPT (Wang et al., 2024). The experimental results in Figure 1 (b) show that these methods failed to capture logical dependencies in 60%, 78%, and 48% of cases, respectively.

Considering the limitations of the existing paradigm and the importance of logical dependencies in fact-checking, we propose a content–logic coupled fact-checking paradigm, which models factuality along two complementary dimensions: content factuality and logical factuality. This paradigm assesses not only the correctness of atomic claims but also the validity of the logical dependencies among them. However, integrating the logical dimension introduces three major challenges:

- **Challenge 1:** Existing fact-checking datasets primarily focus on single facts (Lin et al., 2022;

[1] https://www.politifact.com/

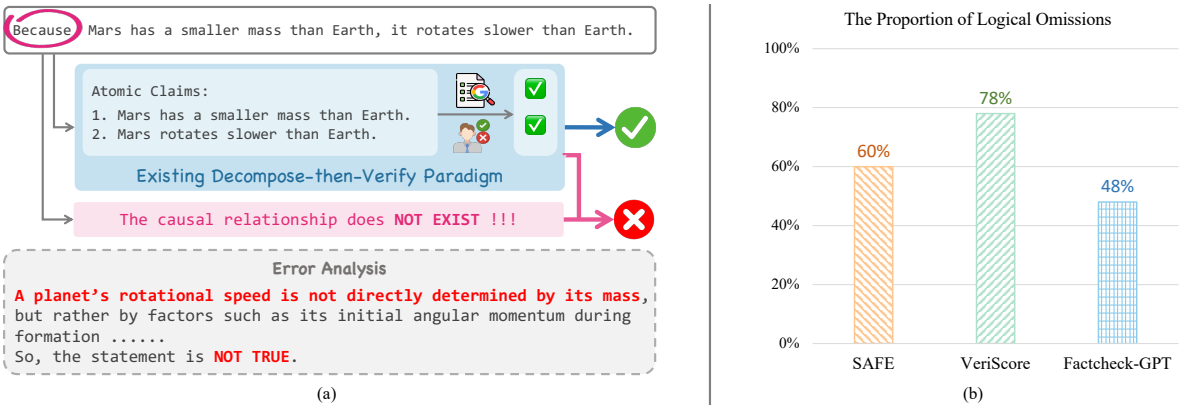

Figure 1: (a) An example illustrating the limitation of the decompose-then-verify paradigm. The text connects two correct facts through an incorrect logical dependency. Existing fact-checking methods verify each atomic claim individually but overlook whether the logical dependency holds, leading to misjudgment of the text as factual. (b) Proportion of logical omissions for three representative fact-checking methods in the preliminary study.

Li et al., 2023). Even more challenging long-form datasets are limited to simple enumerations of facts without modeling relations among them (Min et al., 2023; Wei et al., 2024; Fatahi Bayat et al., 2025). Moreover, existing annotations evaluate factual correctness and coverage by providing reference facts, lacking annotations for logical dependencies and their validity (Chen et al., 2023; Wang et al., 2024; Liu et al., 2025).

- **Challenge 2:** Current fact-checking methods predominantly follow the decompose-then-verify paradigm, emphasizing the completeness and correctness of atomic claim extraction and verification while overlooking logical factuality (Min et al., 2023; Song et al., 2024; Wei et al., 2024; Fatahi Bayat et al., 2025; Liu et al., 2025).
- **Challenge 3:** Traditional metrics such as factual precision and recall (Liu et al., 2025; Wei et al., 2024) measure only fact-level correctness and coverage, failing to capture logical validity.

To address Challenge 1, we introduce a logic-driven fact-checking dataset, LoReFact. It is constructed through a controlled data generation and annotation pipeline that explicitly embeds diverse logical relations into long-form responses and annotates their validity. The dataset contains 1,080 samples with 5,706 annotated logical statements, covering 36 topics across 4 domains. To the best of our knowledge, LoReFact is the first long-form fact-checking dataset that systematically incorporates the logical dimension, as shown in Table 1.

To address Challenge 2, we propose LoReFactcheck, a simple yet effective framework for content–logic coupled factuality evaluation. It comprises two complementary pipelines: content factuality evaluation, which follows standard decompose-and-verify procedures to assess the correctness of atomic claims, and logic factuality evaluation, which detects logical relations, identifies their types, and verifies their validity using external evidence and a multi-model LLM committee.

Benchmark	Logic Annotations	#Prompts	#Topics
FELM	✗	847	5
FActScore	✗	500	1
LongFact	✗	2,280	38
FactCheckBench	✗	94	–
FactRBench	✗	1,096	–
LoReFact (Ours)	✓	1,080	36

Table 1: Comparisions of existing long-form fact-checking benchmarks.

To address Challenge 3, we design a logic-aware factuality metric, LoReFactScore. It combines ContentScore for content correctness and LogicScore for logical validity into a unified score, thereby more faithfully reflecting textual factuality. By explicitly incorporating logical factuality into scoring, LoReFactScore enables the detection and penalization of logical fallacies, while supporting error source decomposition and interpretability.

To summarize: ❶ We introduce a content–logic coupled paradigm for fact-checking, which explicitly incorporates the logical dimension of factuality. ❷ We present a holistic solution consisting of a new long-form fact-checking dataset called LoReFact, an evaluation framework named LoRe-Factcheck, and the LoReFactScore metric, enabling joint evaluation of content and logical factuality. ❸ Experimental results highlight the crucial role of the logical dimension in fact-checking and demonstrate that our proposed paradigm enables a more comprehensive and faithful evaluation of textual factuality.

2 LoReFact

Our current understanding of textual factuality remains limited, as existing fact-checking datasets fail to capture logical dependencies between facts. To fill this gap, our goal is to construct a fact-checking dataset that explicitly incorporates the logical dimension, aiming to obtain texts containing diverse and complex logical statements along with their corresponding logical annotations.

2.1 Data Collection

Prompt Collection The first step of constructing LoReFact is to collect a diverse set of logic-oriented prompts. We manually select 36 topics covering four major domains—engineering and technology, humanities, natural sciences, and social sciences—as illustrated in Figure 2. We then adopt a hybrid strategy that combines manual design with LLM-based generation. For each topic, a small number of logic-driven seed prompts are manually drafted and used as few-shot demonstrations with GPT-4o to generate candidate prompts. After filtering out semantically redundant or logically weak instances, we retain 30 high-quality prompts per topic, resulting in a total of 1,080 prompts.

Response Generation Following the prompt collection, we employ GPT-4o to generate responses for the collected prompts. We design a controlled three-stage process to ensure that logical relations are explicitly and verifiable, rather than implicitly embedded in vague expressions.

Stage 1: Claim Generation. For each prompt, GPT-4o first generates 20 semantically complete claims related to the prompt (10 correct and 10 incorrect). These claims serve as the basic units for subsequent logical composition, enabling diverse error types.

Stage 2: Logical Statement Generation. We follow studies in the field of formal semantics (Asher and Lascarides, 2005; Kamp and Reyle, 1993) and predefine four types of logical relations: causal, conditional, temporal, and concessive. Then, we instruct GPT-4o to generate at least two logical statements for each prompt, based on the predefined logical relations and the generated claims. Each logical statement is required to contain a logical error, either arising from an incorrect logical relation between correct claims or involving at least one incorrect claim. This design reflects common error patterns in real-world logical reasoning, where fallacious statements often stem from flawed inference over accurate information or from reasoning based on false content, thereby resulting in compound errors at both the content and logic levels.

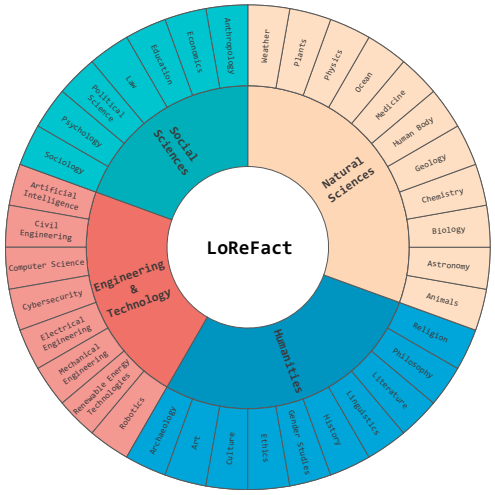

Figure 2: Overview of the four domains and their corresponding 36 topics covered in the LoReFact.

Stage 3: Response Generation. Finally, we employ GPT-4o to draft a response for each prompt based on the generated logical statements. Each response is required to naturally incorporate all given logical statements. Compared with directly generating responses, this process preserves naturalness while significantly improving logical density and verifiability. The detailed instructions used in the data collection process are provided in Appendix A.1.1.

2.2 Data Annotation

Given the diversity of topics and the high density of logical statements, annotation requires strong retrieval and tool-use abilities. Therefore, we recruited three graduate students with research backgrounds in fact-checking and prior annotation experience to annotate the dataset.

The annotation scope covers all logical statements appearing in the responses, including both the controlled logical statements explicitly embedded during generation and the spontaneous logical expressions generated by the model. Each logical statement is annotated along three dimensions: (1) **Content factuality.** For each logical statement, annotators need to identify all included claims and assign binary factuality labels. (2) **Logical relation type.** For each logical statement, annotators label the main type of logical relation involved, including causal, conditional, temporal, and concessive relations. Specifically, a causal relation describes one claim as the cause or result of another; a conditional relation indicates that the truth of one claim depends on the premise set by another; a tempo-

Statistics	All	Engineering & Technology	Humanities	Natural Sciences	Social Sciences
#Sample	1,080	240	300	330	210
Avg. Length	818.67	850.51	808.73	805.94	816.49
#Logical statement	5,706	1,215	1,501	1,843	1,147
– #TRUE	1,277	215	276	495	291
– #FALSE	4,429	1,000	1,225	1,348	856
Agree rate (%)	92.18	94.73	88.14	93.49	92.68

Table 2: Statistics of LoReFact benchmark. "Avg. Length" is the average number of words of responses. "#TRUE"/"#FALSE" denotes the number of the logic factuality of logical statements labeled as true and false respectively. "Agree rate" is the agreement rate of two annotators during annotation.

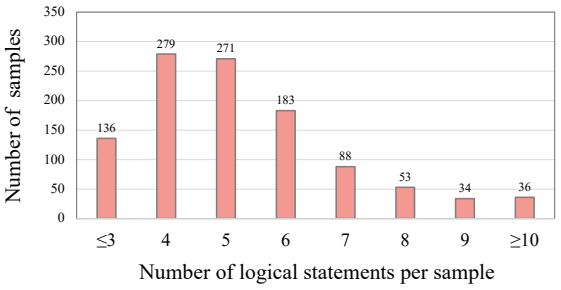

Figure 3: Distribution of samples by the number of logical statements.

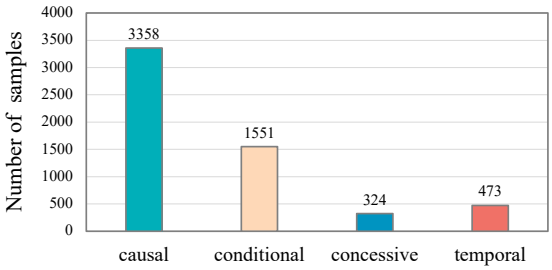

Figure 4: Distribution of samples across different logical relationship types.

ral relation captures the sequence or simultaneity of events among claims; and a concessive relation denotes that one claim remains true despite an apparent conflict with another. (3) **Logical factuality.** Annotators assess whether the claimed logical dependency holds and assign a binary factuality label.

To ensure annotation quality, each sample was independently annotated by two expert annotators, with disagreements resolved by a third annotator to reach consensus. The agreement rate between the two annotators is reported in Table 2, with consensus reached in 92.18% of the cases.

2.3 Data Analysis

Data statistics. Table 2 presents an overview of the LoReFact dataset. It contains 1,080 samples, with an average response length of 818.67 words. It provides annotations for 5,706 logical statements, of which 4,429 contain logical factuality errors. The relatively high proportion of logically false statements stems from our data generation design. As described in Section 2.1, during the logical statement generation stage, we explicitly instruct the model to generate logically flawed statements to ensure sufficient coverage of logical fallacies.

Data characteristics. From the perspective of prompt design, unlike prompts in prior fact-checking datasets that focus on single knowledge points or surface-level factual descriptions, the prompts in LoReFact are logic-oriented, emphasizing logical reasoning and multi-dimensional relational understanding. Specifically, they encourage responses to explain not only what but also why, and often involve interactions among multiple variables in complex systems, naturally leading to associations among multiple factors. We provide illustrative examples in Appendix A.1.2.

From the perspective of response content, each response in LoReFact contains a sufficient number of verifiable logical statements. As shown in Figure 3, most samples contain 3–7 logical statements, balancing logical density and complexity.

Furthermore, we analyze the distribution of logical relation types among these logical statements, as shown in Figure 4. Causal and conditional relations account for the majority, whereas temporal and concessive relations constitute a smaller proportion, consistent with their relative prevalence in natural language narratives. Overall, LoReFact preserves diverse logic types while highlighting common logical patterns.

3 Methodology

3.1 Task Definition and Framework Overview

In our study, fact-checking is formulated as a logic-aware factuality evaluation task. Given a text T, the objective is to compute its overall factuality score using retrieved evidence, detect statements involving logical dependencies, identify the logical relation type, and verify their validity.

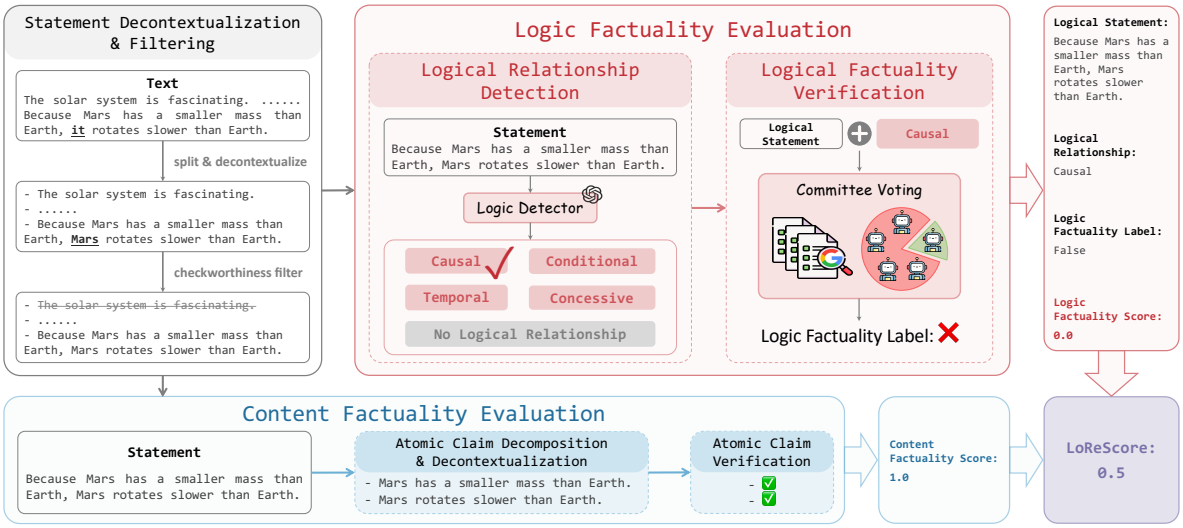

Figure 5: The overall framework of LoRe-Factcheck.

Figure 5 presents an overview of the framework. LoRe-Factcheck first decomposes the input text into multiple statements and then initiates two parallel evaluation pipelines. In the content factuality evaluation, we assess the truthfulness of each statement at the content level and obtain its content factuality score. In the logic factuality evaluation, we examine the correctness of logical dependencies and obtain the corresponding logic factuality score. Finally, LoRe-Factcheck integrates the results from both dimensions to produce a factuality score for each statement and further aggregates them into the overall factuality metric LoReFactScore(T).

3.2 Statement Decomposition and Decontextualization

The text is first split into individual sentences. We then apply decontextualization to resolve contextual dependencies (Choi et al., 2021), making each sentence self-contained (e.g., "it rotates" → "Mars rotates"). Checkworthiness filtering is further performed to exclude non-verifiable sentences, including opinions, questions, or those with little substantive content (e.g., "The solar system is fascinating."). All these steps are conducted by GPT-4o, with further details provided in Appendix A.2.2. The remaining objective and verifiable factual statements serve as the common input for both content and logic factuality evaluation, formally denoted as $S = \{s_1, s_2, ..., s_n\}$.

3.3 Logic Factuality Evaluation

The logic factuality evaluation is devised to assess the validity of the logical dependencies within the statement. This process consists of logical relationship detection and logical factuality verification.

Logical Relationship Detection. We first identify the dominant logical relation type of each statement s_i. This task is framed as a five-class classification problem with the following label set: $\mathcal{T} = \{causal, conditional, temporal, concessive, None\}$. We employ GPT-4o as the logic detector to assign a logical relation type label $\hat{t}_i \in \mathcal{T}$ for each statement, where $\hat{t}_i = None$ indicates that the statement does not involve any logical relation.

Logical Factuality Verification. If $\hat{t}_i = None$, the statement contains no logical dependency and requires no logical factuality verification or scoring. If $\hat{t}_i \neq None$, we further evaluate whether the claimed logical dependency is factually valid. To ensure reliable verification, we first generate a query based on s_i and $\hat{t}_i$ to retrieve the top-3 most relevant evidence from Google Search[2], and then employ a multi-model committee composed of five advanced large language models to assess the logical factuality based on the retrieved evidence. Each model independently outputs a binary judgment $\hat{l}_{i,j}^{\text{logic}} \in \{\text{TRUE}, \text{FALSE}\}$, and the final logical factuality label $\hat{l}_i^{\text{logic}}$ is determined by an unweighted majority vote.

For statements with $\hat{t}_i \neq None$, the logical factuality score LogicScore is defined as follows:

$$\text{LogicScore}(s_i) = \begin{cases} 1, & \text{if } \hat{l}_i^{\text{logic}} = \text{TRUE}, \\ 0, & \text{if } \hat{l}_i^{\text{logic}} = \text{FALSE}. \end{cases} \quad (1)$$

3.4 Content Factuality Evaluation

The content factuality evaluation follows the decompose-then-verify pipeline, aiming to verify

[2]We use Serper as the Google Search API, available at https://serper.dev/.

the factuality of atomic claims within a statement.

Atomic Claim Decomposition and Decontextualization. We first decompose each statement s_i into atomic claims, each containing only one piece of information. Following the same principles as those described in Section 3.2, we decontextualize atomic claims to ensure independent verifiability and filter out subjective opinions. Finally, we obtain the set of atomic claims to be verified for each statement s_i, denoted as $s_i = \{c_{i,1}, c_{i,2}, \ldots, c_{i,m}\}$.

Atomic Claim Verification. We formulate the factual verification of atomic claims as a binary classification task. For each atomic claim $c_{i,k}$, we use it as a query to retrieve the top-3 most relevant document snippets from Google Search as evidence. GPT-4o then serves as the verifier, taking $(c_{i,k}, \text{evidence})$ as input and outputting a binary factual label $\hat{l}_{i,k}^{\texttt{content}} \in \{\text{TRUE}, \text{FALSE}\}$.

We use factual precision to quantify the content factuality of a statement s_i, and denote it as $\text{ContentScore}(s_i)$. It measures the proportion of true atomic claims within a statement.

3.5 LoReFactScore

We propose a logic-aware factuality metric, LoReFactScore, which unifies the results of logic and content factuality evaluation.

For each statement s_i, when s_i does not involve any logical relation, $\text{LoReFactScore}(s_i)$ degenerates into $\text{ContentScore}(s_i)$, which is equivalent to the factual precision used in previous studies. When a logical relation is involved, $\text{LoReFactScore}(s_i)$ incorporates the logical factuality dimension, thereby providing a unified characterization of both content truthfulness and logical validity.

$$\text{LoReFactScore}(s_i) = \begin{cases} \text{C-Score}(s_i), & \text{if } \hat{t}_i = \textit{None}, \\ \alpha \times \text{C-Score}(s_i) + \\ \quad \beta \times \text{L-Score}(s_i), & \text{otherwise}, \end{cases}$$

where C-Score and L-Score denote ContentScore and LogicScore, respectively, and $\alpha, \beta \in [0, 1]$ are trade-off hyperparameters satisfying $\alpha + \beta = 1$. Finally, we aggregate $\text{LoReFactScore}(s_i)$ of all statements to obtain the overall factuality score of the input text T, denoted as:

$$\text{LoReFactScore}(T) = \frac{1}{|S|} \sum_{i=1}^{|S|} \text{LoReFactScore}(s_i), \quad (2)$$

where $S = \{s_1, s_2, ..., s_n\}$ is the set of factual statements in the text. Higher values indicate better overall factuality of the text.

4 Experiments

4.1 Baseline Systems

We choose three representative fact-checking frameworks, including SAFE (Wei et al., 2024), VeriScore (Song et al., 2024), and FactCheck-GPT (Wang et al., 2024), as baselines for comparison. SAFE is the first method in long-form fact-checking that introduces the use of external evidence retrieved from search engines (e.g., Google Search) for verification. VeriScore places greater emphasis on claim verifiability. FactCheck-GPT focuses on fine-grained factuality assessment.

4.2 Evaluation Metrics

We adopt the logic-aware factuality metric $\text{LoReFactScore}(T)$ defined in Section 3.5 as the overall metric, where we set $\alpha = \beta = 0.5$. In addition, we further introduce three logic-related metrics to demonstrate the effectiveness of the proposed framework in evaluating logical factuality:

Logic Recall measures the ability to identify statements that involve logical relations.

$$\text{LogicRecall} = \frac{\sum_{i=1}^{N} \mathbb{1}\left[\hat{t}_i \neq \textit{None} \wedge t_i \neq \textit{None}\right]}{\sum_{i=1}^{N} \mathbb{1}\left[t_i \neq \textit{None}\right]}, \quad (3)$$

where t_i denotes the ground-truth logical type. $\mathbb{1}[\cdot]$ is an indicator function.

Logic Type Accuracy measures accuracy in classifying logical relation types.

$$\text{LogicTypeAcc} = \frac{1}{|S|} \sum_{i=1}^{|S|} \mathbb{1}[\hat{t}_i = t_i]. \quad (4)$$

Logic Factuality Accuracy measures accuracy in assessing the factuality of logical relations.

$$\mathbb{1}_i^{\texttt{logic}} = \mathbb{1}\left[\hat{t}_i = t_i \wedge \hat{l}_i^{\texttt{logic}} = l_i^{\texttt{logic}}\right], \quad (5)$$

$$\text{LogicFactualityAcc} = \frac{\sum_{i=1}^{N} \mathbb{1}_i^{\texttt{logic}}}{\sum_{i=1}^{N} \mathbb{1}\left[\hat{t}_i = t_i\right]}, \quad (6)$$

where $l_i^{\texttt{logic}}$ is the ground-truth logical factuality.

4.3 Main Results

Table 3 shows the comparison results between our LoRe-Factcheck and three representative baselines. We summarize observations below:

LoRe-Factcheck yields stricter yet more faithful overall score. In terms of the overall LoReFactScore, our LoRe-Factcheck framework

Domain	Method	LoReFactScore	Logic Recall	Logic Type Acc	Logic Factuality Acc
All	SAFE	93.01	3.15	–	–
	VeriScore	93.33	1.01	–	–
	Factcheck-GPT	89.23	0.96	–	–
	LoRe-Factcheck (Ours)	59.24 (32.62 ↓)	80.49	78.71	73.70
Engineering & Technology	SAFE	92.36	3.13	–	–
	VeriScore	93.23	1.57	–	–
	Factcheck-GPT	89.59	1.04	–	–
	LoRe-Factcheck (Ours)	57.74 (33.99 ↓)	87.82	79.57	76.38
Humanities	SAFE	93.15	3.16	–	–
	VeriScore	94.08	0.79	–	–
	Factcheck-GPT	87.93	0.99	–	–
	LoRe-Factcheck (Ours)	52.59 (39.13 ↓)	84.68	79.70	72.46
Natural Sciences	SAFE	92.51	2.98	–	–
	VeriScore	92.59	0.83	–	–
	Factcheck-GPT	89.76	0.99	–	–
	LoRe-Factcheck (Ours)	66.65 (24.97 ↓)	73.96	78.80	74.03
Social Sciences	SAFE	94.35	3.44	–	–
	VeriScore	94.03	1.06	–	–
	Factcheck-GPT	89.84	0.79	–	–
	LoRe-Factcheck (Ours)	58.80 (33.94 ↓)	77.77	76.12	71.75

Table 3: Comparison results of different methods on the LOREFACT dataset. Values in parentheses indicate LoRe-Factcheck's score decrease relative to the average of the baselines.

shows an average decrease of around 30 points compared with previous methods. This gap stems from LoRe-Factcheck's explicit penalization of logical-relation errors, yielding a stricter and more faithful assessment of factuality. It further suggests that existing methods tend to underweight logical soundness and thus systematically overestimate overall factuality.

Reliable logic evaluation with robust cross-domain performance. Overall, the proposed LoRe-Factcheck framework performs strongly on Logic Recall, Logic Type Accuracy, and Logic Factuality Accuracy, indicating that it effectively identifies statements involving logical relations. Moreover, across domains, these logic-related metrics exhibit small variance and consistent trends, demonstrating LoRe-Factcheck's strong adaptability and generalization to diverse scenarios with logical dependencies.

Interplay between Logic Recall and Overall Score. We observe that as the framework more aggressively identifies logical relations (higher Logic Recall), a larger portion of statements undergoes logical verification, which exposes and penalizes more potential logical errors—thereby lowering the LoReFactScore. This implies that reporting LoReFactScore alone may yield superficially high scores due to missed logic. LoReFactScore should be interpreted jointly with Logic Recall to distinguish truly more factual outputs from those scoring highly due to insufficient detection.

Domain	Method	LoReFactScore	Logic Recall
All	LoRe-Factcheck	59.24	80.49
	- w/o L-Eval	88.97	1.05
Engineering & Technology	LoRe-Factcheck	57.74	87.82
	- w/o L-Eval	89.11	1.23
Humanities	LoRe-Factcheck	52.59	84.68
	- w/o L-Eval	87.22	0.87
Natural Sciences	LoRe-Factcheck	66.65	73.96
	- w/o L-Eval	89.87	1.19
Social Sciences	LoRe-Factcheck	58.80	77.77
	- w/o L-Eval	89.88	0.87

Table 4: Ablation results on removing Logic Factuality Evaluation (L-Eval) from LoRe-Factcheck.

4.4 Ablation Studies

To further validate the contribution of our logical factuality evaluation module, we design two ablation studies: (i) removing the logic evaluation module and retaining only content-level evaluation; and (ii) replacing the logical-relation detector with different LLMs to examine how detector choice impacts overall performance.

Effect of Removing Logic Factuality Evaluation Module. As shown in Table 4, removing the logic factuality evaluation module raises the overall LoReFactScore from about 60 to 89, which is comparable to the three baselines in Table 3, but results in a marked drop in Logic Recall. This indicates that our content factuality evaluation module is inherently strong, while the lower overall score of the full framework primarily reflects its stricter logical verification. The decline in Logic Recall further underscores that our framework demonstrates solid

Logic Detector	Logic Recall	Logic Type Acc
GPT-4o	80.49	78.71
Gemini-2.5-Pro	78.72	77.76
Deepseek-R1	78.88	74.22
Claude-Opus-4	74.88	74.33
Qwen3-235B-Instruct	80.54	73.26

Table 5: Results of different LLMs as logic detectors.

Method	LoReFactScore	Logic Recall
SAFE	58.37	40.00
VeriScore	60.82	22.00
Factcheck-GPT	57.73	52.00
LoRe-Factcheck (Ours)	40.48	94.00

Table 6: Results of different methods on the preliminary data.

performance in detecting and verifying statements involving logical relations.

Effect of Different Logic Detectors. We further investigate the performance of different LLMs as logical-relation detectors, and the results are shown in Table 5. Overall, models exhibit small gaps in Logic Recall but notably larger differences in Logic Type Accuracy. GPT-4o maintains the highest type classification accuracy while preserving high recall, yielding the most stable overall behavior. Leveraging this advantage, we adopt GPT-4o as the logical-relation detector in our LoRe-Factcheck framework to ensure the reliability of the logical relationship detection module.

4.5 In-depth Analysis of Preliminary Study

Initially, as shown in Figure 1(b), we conducted a preliminary study on 50 real-world political statements to assess whether current fact-checking frameworks overlook logical relations. To further evaluate the effectiveness of our framework on real-world data, we compared it against three baselines. Table 6 reports findings consistent with those on our synthetic LOREFACT dataset. The results suggest that existing methods often neglect logical factuality and consequently overestimate overall factuality, whereas our framework exposes these inconsistencies and yields a more faithful assessment. Moreover, this indicate that explicitly incorporating logical validity into factuality evaluation is necessary in real-world settings.

4.6 Case Study

To better demonstrate the advantage of our content–logic coupled factuality evaluation paradigm, we present a test example along with predictions from different methods in Table 7. In this case, although the two atomic statements are both true

Text	..... Since echolocation involves emitting sound waves and interpreting the echoes that bounce off objects, bats use echolocation mainly at night. .....
Ground Truth	"content factuality": { "atomic claims": ["echolocation involves emitting sound waves and interpreting the echoes that bounce off objects.", "bats use echolocation mainly at night."] "labels": [True, True]} "logical relationship type": "causal" "logical factuality": False
SAFE	"atomic claims": ["echolocation involves emitting sound waves.", "echolocation involves interpreting the echoes that bounce off objects.", "echolocation involves emitting sound waves and interpreting the echoes that bounce off objects.", "bats use echolocation mainly at night."] "verification results": [True, True, True, True] **→ Factual Precision = 4/4 = 1.0**
VeriScore	"atomic claims": ["echolocation involves emitting sound waves and interpreting the echoes that bounce off objects.", "bats use echolocation mainly at night."] "verification results": [True, True] **→ Factual Precision = 2/2 = 1.0**
Factcheck-GPT	"atomic claims": ["echolocation involves emitting sound waves and interpreting the echoes that bounce off objects.", "bats use echolocation mainly at night."] "verification results": [True, True] **→ Factual Precision = 2/2 = 1.0**
LoRe-Factcheck (Ours)	"atomic claims": ["echolocation involves emitting sound waves and interpreting the echoes that bounce off objects.", "bats use echolocation mainly at night."] "verification results": [True, True] → ContentScore = Factual Precision = 2/2 = 1.0 "logical relationship type": "causal" "logical verification result": False → LogicScore = 0.0 **→ LoReFactScore = ContentScore × 0.5 + LogicScore × 0.5 = 0.5**

Table 7: Prediction comparison between different methods on a test sample.

at the content level, the causal relation asserted between them does not hold. The three representative baselines assess only content factuality and therefore assign a perfect Factual Precision (i.e., our ContentScore) = 1.0, failing to detect the logical error. In contrast, our LoRe-Factcheck detects the logical relation type (causal) and further determines its logical factuality to be False. Moreover, under our overall (LoReFactScore) metric, the case receives a factuality score of 0.5. This reduction indicates that our paradigm exposes and penalizes hidden logical fallacies, thereby providing a more comprehensive and reliable factuality assessment.

5 Conclusion

We propose a content–logic coupled evaluation paradigm for fact-checking. Following this paradigm, we provide a complete set of resources consisting of a logic-driven long-form fact-checking dataset, a joint content–logic evaluation framework, and a logic-aware metric for comprehensive factuality assessment. Experimental results demonstrate that our proposed paradigm leads to more reliable factuality evaluation. We hope this work provides useful foundations and resources for future studies on logic-aware fact-checking.

Limitations

While our work introduces a content-logic coupled evaluation paradigm to bridge the logic gap in fact-checking, there are still several limitations remain. (1) Although we focus on fundamental logical relations commonly found in narrative texts, these relations do not encompass all complex reasoning structures. (2) LoReFact is constructed through controlled LLM generation, which ensures high logical density and annotation reliability, but may not fully capture the complexity of logical fallacies in natural texts and may exhibit distributional shift. (3) LoRe-Factcheck does not guarantee global consistency across multi-step reasoning chains, as our work focuses on the logical dimension systematically overlooked in fact-checking rather than complete reasoning verification.

Future work can mitigate these limitations by expanding the real-world data from our preliminary study, incorporating broader logical fallacy types, and exploring methods to verify global logical consistency across reasoning chains.

Ethical Considerations

During the dataset construction process, all content generated by LLMs was manually reviewed by the authors to ensure that the LLM-generated questions and responses included in the dataset do not pose any ethical concerns or introduce undesirable biases. In addition, all annotators involved in the data annotation process were paid at rates above the local minimum wage.

References

Anthropic. 2025. Introducing claude 4.

Nicholas Asher and Alex Lascarides. 2005. *Logics of Conversation*. Studies in natural language processing. Cambridge University Press.

Shiqi Chen, Yiran Zhao, Jinghan Zhang, I-Chun Chern, Siyang Gao, Pengfei Liu, and Junxian He. 2023. FELM: benchmarking factuality evaluation of large language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Qinyuan Cheng, Tianxiang Sun, Wenwei Zhang, Siyin Wang, Xiangyang Liu, Mozhi Zhang, Junliang He, Mianqiu Huang, Zhangyue Yin, Kai Chen, and Xipeng Qiu. 2023. Evaluating hallucinations in chinese large language models. *CoRR*, abs/2310.03368.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. 2023. Factool: Factuality detection in generative AI - A tool augmented framework for multi-task and multi-domain scenarios. *CoRR*, abs/2307.13528.

Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. 2021. Decontextualization: Making sentences stand-alone. *Transactions of the Association for Computational Linguistics*, 9:447–461.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit S. Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 81 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *CoRR*, abs/2507.06261.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 81 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

Farima Fatahi Bayat, Lechen Zhang, Sheza Munir, and Lu Wang. 2025. FactBench: A dynamic benchmark for in-the-wild language model factuality evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 33090–33110, Vienna, Austria. Association for Computational Linguistics.

Anisha Gunjal and Greg Durrett. 2024. Molecular facts: Desiderata for decontextualization in LLM fact verification. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3751–3768, Miami, Florida, USA. Association for Computational Linguistics.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *CoRR*, abs/2301.07597.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Qisheng Hu, Quanyu Long, and Wenya Wang. 2025. Decomposition dilemmas: Does claim decomposition boost or burden fact-checking performance? In *Proceedings of the 2025 Conference of the Nations*

of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 6313–6336, Albuquerque, New Mexico. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2):42:1–42:55.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38.

Zhengping Jiang, Jingyu Zhang, Nathaniel Weir, Seth Ebner, Miriam Wanner, Kate Sanders, Daniel Khashabi, Anqi Liu, and Benjamin Van Durme. 2024. Core: Robust factual precision scoring with informative sub-claim identification. *CoRR*, abs/2407.03572.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. WiCE: Real-world entailment for claims in Wikipedia. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7561–7583, Singapore. Association for Computational Linguistics.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic - Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*, volume 42 of *Studies in linguistics and philosophy*. Springer.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Miaoran Li, Baolin Peng, Michel Galley, Jianfeng Gao, and Zhu Zhang. 2024. Self-checker: Plug-and-play modules for fact-checking with large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 163–181, Mexico City, Mexico. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Xin Liu, Lechen Zhang, Sheza Munir, Yiyang Gu, and Lu Wang. 2025. VeriFact: Enhancing long-form factuality evaluation with refined fact extraction and reference facts. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17919–17936, Suzhou, China. Association for Computational Linguistics.

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

Yixiao Song, Yekyung Kim, and Mohit Iyyer. 2024. VeriScore: Evaluating the factuality of verifiable claims in long-form text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9447–9474, Miami, Florida, USA. Association for Computational Linguistics.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2024. FreshLLMs: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13697–13720, Bangkok, Thailand. Association for Computational Linguistics.

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, Isabelle Augenstein, Iryna Gurevych, and Preslav Nakov. 2024. Factcheck-bench: Fine-grained evaluation benchmark for automatic fact-checkers. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14199–14230, Miami, Florida, USA. Association for Computational Linguistics.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. 2024. Long-form factuality in large language models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 40 others. 2025. Qwen3 technical report. *CoRR*, abs/2505.09388.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the AI ocean: A survey on hallucination in large language models. *CoRR*, abs/2309.01219.

A Appendix

A.1 Details of LoReFact

A.1.1 Data construction process

To ensure broad domain coverage, the topics in LoReFact are manually selected with reference to multiple existing resources, including MMLU (Hendrycks et al., 2021), SAFE (Wei et al., 2024), Quora (quora.com), and HC3 (Guo et al., 2023). The complete list of domains and topics is shown in Table 8.

To guide the generation of logic-oriented questions, we designed a small set of high-quality topic–question pairs as in-context demonstrations. Examples of these demonstrations were shown in Table 9. Each demonstration was manually verified to ensure that the corresponding question encouraged long-form responses involving explicit reasoning structures rather than surface-level factual recall. Prompt templates for question generation is shown in Table 10.

To ensure that the generated responses explicitly contain verifiable logical relations, we employed a controlled three-stage generation pipeline to ensure controllability. Prompt templates for claim generation, logical statement generation, and response generation are shown in Table 11, Table 12, and Table 13, respectively.

A.1.2 Comparison of LoReFact with other benchmarks

As shown in Table 14, we illustrate the design differences between prompts in LoReFact and those in existing long-form fact-checking benchmarks. The examples indicate that prior datasets primarily emphasize listing or summarizing factual information, whereas LoReFact adopts a logic-oriented design that encourages reasoning and generation from a logical perspective.

A.1.3 Examples of LoReFact

We provide an example prompt for each of the topics in LoReFact. These example prompts are shown in Table 15, 16, 17, 18.

A.2 Details of LoRe-Factcheck

A.2.1 Implementation

For the baselines, we follow the official implementations released by the original authors. To ensure a fair comparison, we uniformly employ GPT-4o to perform all LLM-driven operations, including atomic claim decomposition, decontextualization, and verification. Since the baselines do not model logical relations, their scores are entirely determined by the content factuality score. For the logic-related metrics, we only estimate Logic Recall indirectly. Specifically, considering that the process of atomic claim decomposition may also retrieve a subset of logic-related claims, we combine semantic similarity computation with manual reviewing to determine which logic-related claims are covered and thereby estimate recall.

For our proposed LoRe-Factcheck, logical factuality verification is conducted by a model committee composed of GPT-4o, Gemini-2.5-Pro (Comanici et al., 2025), Claude-Opus-4 (Anthropic, 2025), DeepSeek-R1 (DeepSeek-AI et al., 2025), and Qwen3-235B-Instruct (Yang et al., 2025). All remaining steps are performed using GPT-4o.

A.2.2 Instructions used in LoRe-Factcheck

We present all instructions used in LoRe-Factcheck in Table 19, 20, 21, 22, 23, 24.

Decontextualization. To address the issue that some sentences rely on contextual information, we perform sentence-level decontextualization by explicitly resolving pronouns and implicit entity mentions, so that each sentence can convey its original meaning independently. The specific instructions for decomposition and decontextualization at the sentence and atomic-claim levels are provided in Table 19 and Table 23.

Checkworthiness Filtering. Not all sentences obtained after segmentation and decontextualization are factual statements. Some sentences are not objectively verifiable, including opinions, questions, exclamations, imperatives, or sentences with little substantive information (e.g., introductory or closing greetings). We therefore apply a checkworthiness filtering step to remove such non-verifiable cases. The same instruction is used for checkworthiness filtering at both the sentence and atomic-claim levels, as provided in Table 20.

B Related Work

B.1 Decompose-then-Verify Paradigm

With the rapid growth of online information, effective fact-checking has become crucial to prevent the spread of misinformation. Given that the text often conveys multiple pieces of information, evaluating its factuality as a whole is challenging (Li et al., 2023). Existing research mainly follows the

decompose-then-verify paradigm. Following this paradigm, early research primarily focused on how to effectively decompose input text into atomic claims (Min et al., 2023; Kamoi et al., 2023). Subsequent work further investigated the verifiability of these atomic claims (Jiang et al., 2024; Song et al., 2024; Fatahi Bayat et al., 2025). Meanwhile, other studies revealed limitations of existing approaches, such as challenges introduced during decomposition and decontextualization, showing that improper claim splitting may distort semantics or omit critical information (Gunjal and Durrett, 2024; Hu et al., 2025). More recent work has also recognized that facts in long-form text are not independent but often exhibit dependencies, and has attempted to address information omission through more fine-grained fact extraction, thereby improving recall and accuracy in fact verification (Liu et al., 2025). In addition, some studies leverage large language models as agents and combine them with search engines (e.g., Google Search) to perform retrieval and verification (Chern et al., 2023; Wei et al., 2024; Li et al., 2024). While these methods enhance fact extraction and external evidence acquisition, their primary focus remains on improving factual coverage and verification, thereby overlooking the validity of logical dependencies among facts. As a result, texts containing logical fallacies may still be misjudged as factual.

In contrast, our proposed LoRe-FactCheck framework extends the conventional decompose-then-verify pipeline by explicitly detecting and validating logical relationships among facts, enabling content–logic coupled factuality evaluation.

B.2 Fact-checking Datasets and Metrics

Early datasets (such as TruthfulQA (Lin et al., 2022), HaluEval (Li et al., 2023), FreshQA (Vu et al., 2024), and HalluQA (Cheng et al., 2023)) mainly focus on single factual knowledge. Even current more challenging long-form fact-checking datasets (FActScore (Min et al., 2023), FELM (Chen et al., 2023), LongFact (Wei et al., 2024), and FactBench (Fatahi Bayat et al., 2025)) consist of biographical questions or simple declarative questions. Their responses are mainly linear narratives that only need to list facts. Moreover, existing datasets mainly provide annotations for factual correctness and coverage based on reference fact sets, without explicitly annotating the logical relationships among facts and their validity (Chen et al., 2023; Wang et al., 2024; Liu et al., 2025). In contrast, LoReFact adopts a logic-driven dataset design, providing explicit annotations of logical relationships among facts and their validity, thereby filling the gap in logical factuality that has been largely overlooked in existing fact-checking datasets.

Regarding metrics, traditional metrics reflect textual factuality from the perspective of atomic claims through factual precision or recall (Wei et al., 2024; Liu et al., 2025). LoReFactScore is the first structured metric to explicitly incorporate logical factuality into scoring, enabling logical fallacies to be directly exposed and penalized. By decoupling content and logical factuality, it further supports error source decomposition and enhances interpretability.

Existing evaluation datasets, frameworks, and metrics primarily center on atomic claims, while neglecting the logical dimension. Our content–logic coupled factuality evaluation paradigm addresses this limitation by jointly assessing content and logical dimensions, providing a comprehensive and reliable view of factuality.

Topic	Domain
Anthropology	Social Sciences
Archaeology	Humanities
Art	Humanities
Artificial Intelligence	Engineering & Technology
Astronomy	Natural Sciences
Biology	Natural Sciences
Botany	Natural Sciences
Chemistry	Natural Sciences
Civil Engineering	Engineering & Technology
Computer Science	Engineering & Technology
Culture	Humanities
Cybersecurity	Engineering & Technology
Economics	Social Sciences
Education	Social Sciences
Electrical Engineering	Engineering & Technology
Ethics	Humanities
Gender Studies	Humanities
Geology	Natural Sciences
History	Humanities
Human Biology	Natural Sciences
Law	Social Sciences
Linguistics	Humanities
Literature	Humanities
Mechanical Engineering	Engineering & Technology
Medicine	Natural Sciences
Meteorology	Natural Sciences
Oceanography	Natural Sciences
Philosophy	Humanities
Physics	Natural Sciences
Political Science	Social Sciences
Psychology	Social Sciences
Religion	Humanities
Renewable Energy Technologies	Engineering & Technology
Robotics	Engineering & Technology
Sociology	Social Sciences
Zoology	Natural Sciences

Table 8: Summary of all 36 topics used in LOREFACT.

Topic	Questions
Animals	- How have deep-sea organisms adapted to extreme pressures, perpetual darkness, and scarce nutrient sources, and what unique sensory and metabolic strategies (e.g., chemosynthesis, bioluminescence) have emerged in these isolated environments? - Why did the evolution of flight evolve independently in such phylogenetically distant groups as insects, pterosaurs, birds, and bats, and how did the unique structural adaptations (e.g., wing membrane vs. feather) in each lineage constrain their subsequent ecological diversification? - How did continental drift during the Triassic, Jurassic, and Cretaceous periods influence the diversification and geographic distribution of dinosaur species? - Why do some migratory species, like the Arctic tern or monarch butterfly, undertake phenomenally long-distance journeys across generations, and how do they integrate innate genetic programming with environmental cues for navigation and timing? - Why did the evolution of the amniotic egg represent such a pivotal reproductive innovation for vertebrates, and how did it enable the colonization of dry, terrestrial environments by reptiles, birds, and mammals?
Computer Science	- How do operating systems manage system resources such as CPU, memory, and I/O devices to ensure efficient and fair usage by multiple applications? - Why did the von Neumann architecture become the dominant model for general-purpose computers, and how do its inherent limitations drive research into alternative computing paradigms like quantum or neuromorphic computing? - How do compilers translate high-level programming languages into machine code, and what optimizations are commonly used to improve performance? - Why do modern operating systems implement hierarchical memory management with multiple cache levels, and how do these designs mitigate the growing processor-memory performance gap? - How did the separation of concerns in software engineering (e.g., MVC architecture) evolve, and why does this principle remain central to managing complexity in large codebases?
Economics	- What are the key factors that determine the elasticity of demand for essential goods and services, and how do these factors impact pricing strategies? - Why do certain industries tend towards natural monopoly structures, and how do regulatory interventions attempt to balance efficiency gains from scale with the social costs of market power? - Why do financial markets experience cycles of boom and bust, and what role do investor expectations and speculative activities play in these cycles? - Why do some regions within a country experience higher economic growth rates than others, and how do policies like regional development funds address these disparities? - Why did the Bretton Woods system of fixed exchange rates collapse, and how does the modern regime of floating rates interact with national monetary policy autonomy and international capital flows?
Literature	- Why do certain canonical texts sustain multiple, even conflicting interpretations over time, and how does this interpretive openness relate to their internal logical structure? - Why do certain literary movements emerge in response to political or economic crises, and how are these conditions encoded in narrative form rather than explicit themes? - How did social class structures shape character motivation and plot development in nineteenth-century realist novels? - Why did fragmented narrative structures become a dominant strategy for representing trauma in post-war literature? - How do metafictional strategies challenge traditional causal logic between plot, character, and meaning in postmodern literature?

Table 9: We provide a set of twenty manually-created in-context demonstrations across four topics.

Instruction:
1. Given a TOPIC, generate a naturally phrased scientific QUESTION.
2. The question must be open-ended and begin with a question word such as "Why", "How", "What", or "Does".
3. The question should focus on a specific and observable phenomenon or mechanism within the given topic, with a clear and well-defined scope.
4. The question should be designed to elicit answers containing multiple logical relationships, such as causal, conditional, or temporal reasoning, resulting in rich and logically structured explanations.
5. Avoid vague or overly broad questions. Focus on a concrete concept, process, or contradiction that is interesting, counterintuitive, or fundamental.
6. The question must be factually grounded and suitable for an informative scientific or educational answer.
7. Do not include subjective opinions or speculative content in the question.
8. Wrap the question in square brackets.
9. Your task is to do this for the TOPIC under "Your Task". Some examples have been provided for you to learn how to do this task.

TOPIC:
[EXAMPLE TOPIC #1]

QUESTION:
[EXAMPLE QUESTION #1]

.....

TOPIC:
[EXAMPLE TOPIC #N]

QUESTION:
[EXAMPLE QUESTION #N]

Your Task:
TOPIC:
[TOPIC_PLACEHOLDER]

QUESTION:

Table 10: Prompt template used for question generation. The [TOPIC_PLACEHOLDER] will be replaced by a given topic term, and the [EXAMPLE TOPIC #1→N] together with their corresponding [EXAMPLE QUESTION #1→N] are filled with N topic–question pairs serving as in-context exemplars.

Instruction:
1. Generate 20 atomic claims related to the given question, including 10 factually correct claims and 10 factually incorrect claims.
2. An atomic claim is a sentence containing a singular piece of information, which should be a basic, indivisible, and independently verifiable.
3. The content of each atomic claim should be as rich and diverse as possible, covering different aspects of the given QUESTION.
4. Provide the atomic claims and wrap them in a markdown code block with an unordered list format, listing each atomic claim as a separate bullet point (-).
5. Your task is to do this for the QUESTION under "Your Task". Some examples have been provided for you to learn how to do this task.

EXAMPLE #1:
QUESTION:
Does the weight of the planet really impact the speed it would move along its own axis?

ATOMIC CLAIMS:
The correct atomic claims are as follows:
```
- The angular momentum acquired during a planet's formation primarily determines its rotational speed.
- Mars has a smaller mass than Earth.
.....
- Venus rotates extremely slowly.
```
The incorrect atomic claims are as follows:
```
- The Moon's rotational period is not synchronized with Earth's.
- Mars rotates faster than Earth.
.....
- A planet's rotational speed is entirely controlled by the strength of its magnetic field.
```

EXAMPLE #2:
QUESTION:
What makes hurricanes so powerful?

ATOMIC CLAIMS:
The correct atomic claims are as follows:
```
- Hurricanes draw their energy from the warm ocean waters they pass over.
- The Coriolis effect contributes to the rotation of hurricanes.
.....
- A well-organized eye structure usually indicates a powerful hurricane.
```
The incorrect atomic claims are as follows:
```
- The eye of a hurricane is typically the most violent part of the storm.
- Hurricanes can sustain themselves without contact with warm waters.
.....
- The strongest hurricanes occur in polar regions.
```

.....

EXAMPLE #N:
QUESTION:
[EXAMPLE QUESTION #N]

ATOMIC FACTS:
The correct atomic claims are as follows:
```
[EXAMPLE CORRECT ATOMIC FACT #N-#1]
.....
[EXAMPLE CORRECT ATOMIC FACT #N-#M]
```
The incorrect atomic claims are as follows:
```
[EXAMPLE INCORRECT ATOMIC FACT #N-#1]
.....
[EXAMPLE INCORRECT ATOMIC FACT #N-#M]
```

Your Task:
QUESTION:
[QUESTION_PLACEHOLDER]

ATOMIC CLAIMS:

Table 11: Prompt template used for claim generation. The [QUESTION_PLACEHOLDER] will be replaced by a given question. For clarity and conciseness, we present two complete examples. The placeholders [EXAMPLE QUESTION #N] and [EXAMPLE CORRECT/INCORRECT ATOMIC FACT #N-#1→M] are filled with the corresponding questions and atomic claims, serving as in-context exemplars.

Instruction:
1. Given a list of ATOMIC CLAIMS, generate at least 2 LOGICAL STATEMENTS by logically combining two or more atomic claims.
2. The logical relationship between atomic claims should be one of the following types: causal, conditional, temporal, or concessive.
3. DO NOT introduce any additional information that is not contained in the atomic claims, and DO NOT alter the components of the atomic claims in the logical statement.
4. Each logical statement must contain a logical error. Specifically, this falsehood must arise in one of the following ways:
   - An incorrect or illogical relationship connecting atomic claims that are all factually correct, or
   - An incorrect or illogical relationship in which at least one of the involved atomic claims is factually incorrect.
5. Do NOT explicitly indicate uncertainty or falsity (e.g., avoid using words like "assume", "might", etc.).
6. Make the logical statement fluent, coherent, and logically connected on the surface, even with the inclusion of false logic.
7. Provide the logical statements and wrap them in a markdown code block with an unordered list format, listing each logical statement as a separate bullet point (-).
8. Your task is to do this for the ATOMIC CLAIMS under "Your Task". Some examples have been provided for you to learn how to do this task.

EXAMPLE #1:
ATOMIC CLAIMS:
The correct atomic claims are as follows:
```
- The angular momentum acquired during a planet's formation primarily determines its rotational speed.
- Mars has a smaller mass than Earth.
- Changes in mass distribution have a greater impact on a planet's rotational speed than changes in total mass.
- Tidal interactions between Earth and the Moon gradually slow down Earth's rotation.
.....
- Venus rotates extremely slowly.
```
The correct atomic claims are as follows:
```
- The Moon's rotational period is not synchronized with Earth's.
- Mars rotates faster than Earth.
- Earth's rotational speed has remained constant.
- An increase in a planet's mass directly causes a proportional increase in its rotational speed.
.....
- A planet's rotational speed is entirely controlled by the strength of its magnetic field.
```

LOGICAL STATEMENTS:
```
- Because the angular momentum acquired during a planet's formation primarily determines its rotational speed, so Venus rotates extremely slowly. [causal]
- Tidal interactions between Earth and the Moon gradually slow down Earth's rotation after the Moon's rotational period is not synchronized with Earth's. [temporal]
- If changes in mass distribution have a greater impact on a planet's rotational speed than changes in total mass, then Venus rotates extremely slowly. [conditional]
.....
- Although an increase in a planet's mass is often assumed to directly cause a proportional increase in its rotational speed, Mars rotates faster than Earth despite having a smaller mass than Earth. [concessive]
```

.....

EXAMPLE #N:
ATOMIC CLAIMS:
The correct atomic claims are as follows:
```
[EXAMPLE CORRECT ATOMIC FACT #N-#1]
.....
[EXAMPLE CORRECT ATOMIC FACT #N-#M]
```
The incorrect atomic claims are as follows:
```
[EXAMPLE INCORRECT ATOMIC FACT #N-#1]
.....
[EXAMPLE INCORRECT ATOMIC FACT #N-#M]
```

LOGICAL STATEMENTS:
```
[EXAMPLE LOGICAL STATEMENT #N-#1]
.....
[EXAMPLE LOGICAL STATEMENT #N-#M]
```

Your Task:
ATOMIC CLAIMS:
[ATOMIC_CLAIMS_PLACEHOLDER]

LOGICAL STATEMENTS:

Table 12: Prompt template used for logcial statement generation. The [ATOMIC_CLAIMS_PLACEHOLDER] will be replaced by the given atomic claims. For clarity and conciseness, we present one complete example. The placeholders [EXAMPLE CORRECT/INCORRECT ATOMIC FACT #N-#1→M] and [EXAMPLE LOGICAL STATEMENT #N-#1→M] are filled with the corresponding atomic claims and logical statements, serving as in-context exemplars.

Instruction:
1. Given a QUESTION and a list of LOGICAL STATEMENTS, generate a RESPONSE.
2. Ensure the RESPONSE includes all given logical statements, exhibits clear logical organization, and is no fewer than 500 words.
3. The RESPONSE should use precise and formal language, avoiding vague generalizations and rhetorical fillers, with a clear focus on delivering informative content.
4. The output should be wrapped in a code block and contain only the RESPONSE.
5. Your task is to do this for the QUESTION under "Your Task". Some examples have been provided for you to learn how to do this task.

EXAMPLE #1:
QUESTION:
Does the weight of the planet really impact the speed it would move along its own axis?

LOGICAL STATEMENTS:
```
- Because the angular momentum acquired during a planet's formation primarily determines its rotational speed, so Venus rotates extremely slowly.
- Tidal interactions between Earth and the Moon gradually slow down Earth's rotation after the Moon's rotational period is not synchronized with Earth's.
- If changes in mass distribution have a greater impact on a planet's rotational speed than changes in total mass, then Venus rotates extremely slowly.
.....
- Although an increase in a planet's mass is often assumed to directly cause a proportional increase in its rotational speed, Mars rotates faster than Earth despite having a smaller mass than Earth.
```

RESPONSE:
```
The weight—or more precisely, the mass—of a planet does not directly determine how fast it rotates around its own axis. Instead, a planet's rotation speed ..... Because the angular momentum acquired during a planet's formation primarily determines its rotational speed, so Venus rotates extremely slowly. .....
That said, mass can indirectly influence rotation rate through its effects on planetary structure and interactions. Although an increase in a planet's mass is often assumed to directly cause a proportional increase in its rotational speed, Mars rotates faster than Earth despite having a smaller mass than Earth. More massive planets ..... If changes in mass distribution have a greater impact on a planet's rotational speed than changes in total mass, then Venus rotates extremely slowly. Additionally, massive gas giants .....
On the other hand, smaller rocky planets like Mercury ..... Tidal interactions between Earth and the Moon gradually slow down Earth's rotation after the Moon's rotational period is not synchronized with Earth's. .....
In conclusion, ..... The initial conditions of formation and subsequent gravitational interactions have far more direct influence on how fast a planet rotates.
```

.....

EXAMPLE #N:
QUESTION:
[EXAMPLE QUESTION #N]

LOGICAL STATEMENTS:
```
[EXAMPLE LOGICAL STATEMENT #N-1]
.....
[EXAMPLE LOGICAL STATEMENT #N-M]
```

RESPONSE:
```
[EXAMPLE RESPONSE #N]
```

Your Task:
QUESTION:
[QUESTION_PLACEHOLDER]

LOGICAL STATEMENTS:
[LOGICAL_STATEMENTS_PLACEHOLDER]

RESPONSE:

Table 13: Prompt template used for response generation. [QUESTION_PLACEHOLDER] will be replaced by the given question, and [LOGICAL_STATEMENT_PLACEHOLDER] will be replaced by the given logical statements. For clarity and conciseness, we present one complete example. The placeholders [EXAMPLE QUESTION #N], [EXAMPLE LOGICAL STATEMENT #N-#1→M], and [EXAMPLE RESPONSE #N] are filled with the corresponding question, logical statements, and response, serving as in-context exemplars. The locations of the given logical statements within the response are highlighted with a blue background to facilitate readability.

Dataset	Example
FELM	Tell me about the World Happiness Report.
FActScore	Tell me a bio of Bridget Moynahan.
LongFact	Can you provide an overview of the International Monetary Fund?
FactCheckBench	Give me a list of the top 10 tallest buildings in the world.
FactRBench	what states are part of the southeast region of Brazil?
LoReFact (Ours)	How do plants regulate gas exchange through stomata, and what factors influence their opening and closing mechanisms?

Table 14: Comparison of prompts in LoReFact and existing long-form fact-checking benchmarks.

Domain	Topic	Example
Engineering & Technology	Artificial Intelligence	"Why does the choice of activation function impact the performance and behavior of deep learning models?"
	Civil Engineering	"Why do different bridge designs, such as suspension and arch bridges, exhibit varying levels of efficiency in distributing loads and resisting environmental forces?"
	Computer Science	"Why do certain algorithms exhibit exponential time complexity, and what strategies are used to mitigate this in practical applications?"
	Cybersecurity	"Why do certain types of malware, such as ransomware, exhibit rapid adaptation and evolution in response to cybersecurity defenses?"
	Electrical Engineering	"Why does the frequency of alternating current influence the operation of transformers and electrical motors?"
	Mechanical Engineering	"How do material properties such as strength, ductility, and thermal conductivity influence the design of mechanical components in engineering systems?"
	Renewable Energy Technologies	"How do advancements in photovoltaic cell materials improve the efficiency and cost-effectiveness of solar energy systems?"
	Robotics	"How do advancements in sensor technology improve the ability of robots to perceive and interact with their environments?"

Table 15: Topics and examples from the Engineering & Technology domain in the LoReFact.

Domain	Topic	Example
Humanities	Archaeology	"What can isotope analysis of human remains reveal about migration patterns and diet in ancient populations?"
	Art	"How does the interaction between light and different types of varnishes used in paintings alter the visual perception of color and texture?"
	Culture	"How does the concept of cultural relativism help explain the diversity of moral and ethical systems across societies?"
	Ethics	"What mechanisms determine how individuals prioritize competing ethical obligations in complex situations?"
	Gender Studies	"What impact does the inclusion of non-binary and transgender identities have on traditional frameworks of gender studies?"
	History	"What impact does the inclusion of non-binary and transgender identities have on traditional frameworks of gender studies?"
	Linguistics	"How does the study of syntax help us understand the underlying rules and structures that govern sentence formation in different languages?"
	Literature	"How do narrative structures in literature influence the reader's perception of time and causality within a story?"
	Philosophy	"Why do different philosophical traditions interpret the nature of consciousness in fundamentally distinct ways?"
	Religion	"Why do different philosophical traditions interpret the nature of consciousness in fundamentally distinct ways?"

Table 16: Topics and examples from the Humanities domain in the LoReFact.

Domain	Topic	Example
Natural Sciences	Animals	"How do animals with specialized coloration, such as camouflage or warning coloration, use these adaptations to survive in their specific environments?"
	Astronomy	"How do gravitational interactions between celestial bodies influence the formation and stability of planetary systems over time?"
	Biology	"Why do certain organisms exhibit cooperative behavior in ecosystems, and what evolutionary mechanisms drive this phenomenon?"
	Chemistry	"Why does the rate of a chemical reaction change with temperature, and how does the Arrhenius equation mathematically describe this relationship?"
	Geology	"Why do certain minerals crystallize in specific shapes, and how does their atomic structure influence their physical properties?"
	Human Body	"How does the regulation of blood glucose levels involve coordinated feedback mechanisms between the pancreas, liver, and other tissues?",
	Medicine	"How does the regulation of blood glucose levels involve coordinated feedback mechanisms between the pancreas, liver, and other tissues?",
	Ocean	"Why does the thermohaline circulation play a critical role in regulating Earth's climate, and what mechanisms drive its movement?"
	Physics	"Why do gyroscopic effects stabilize rotating objects, and what physical principles govern their precession under external forces?"
	Plants	"How do plants regulate gas exchange through stomata, and what factors influence their opening and closing mechanisms?"
	Weather	"How do variations in atmospheric pressure drive the formation and movement of high- and low-pressure systems, influencing weather patterns globally?"

Table 17: Topics and examples from the Natural Sciences domain in the LOREFACT.

Domain	Topic	Example
Social Sciences	Anthropology	"Why did bipedalism evolve in early hominins, and what environmental or anatomical factors may have driven this transition?"
	Economics	"Why do some economies experience hyperinflation, and what mechanisms drive this extreme devaluation of currency?"
	Education	"Why do some teaching methods, such as inquiry-based learning, foster deeper conceptual understanding compared to traditional lecture-based approaches?"
	Law	"How does the principle of stare decisis influence the consistency and predictability of judicial decisions in common law systems?"
	Political Science	"How does the distribution of economic resources within a society influence patterns of political participation and voter turnout?"
	Psychology	"How does the distribution of economic resources within a society influence patterns of political participation and voter turnout?"
	Sociology	"Why do individuals conform to group norms, and what mechanisms drive this conformity in different social settings?"

Table 18: Topics and examples from the Social Sciences domain in the LOREFACT.

Your task is to perform sentence segmentation and de-contextualization.
Let's define a function named process(input:str).
The return value should be a list of strings, where each string should be a decontextualized sentence.
For example, if a user call process("Mary is a five-year old girl. She likes playing piano. She doesn't like cookies.").
You should return a python list without any other words,
["Mary is a five-year old girl.", "Mary likes playing piano.", "Mary doesn't like cookies."]
Note that your response will be passed to the python interpreter, SO NO OTHER WORDS!

process([TEXT_PLACEHOLDER])

Table 19: Instruction used for statement decomposition and decontextualization. The [TEXT_PLACEHOLDER] will be replaced by a given text.

Your task is to identify whether texts are checkworthy in the context of fact-checking.
Let's define a function named checkworthy(input: List[str]).
The return value should be a list of strings, where each string selects from ["Yes", "No"].
"Yes" means the text is a factual checkworthy statement.
"No" means that the text is not checkworthy, it might be an opinion, a question, or others.

For example, if a user call checkworthy(["I think Apple is a good company.", "Friends is a great TV series.", "Are you sure Preslav is a professor in MBZUAI?", "The Stanford Prison Experiment was conducted in the basement of Encina Hall.", "As a language model, I can't provide these info."]).
You should return a python list without any other words,
["No", "Yes", "No", "Yes", "No"]

Note that your response will be passed to the python interpreter, SO NO OTHER WORDS!

checkworthy([TEXTS_PLACEHOLDER])

Table 20: Instruction used for statement and atomic claim filtering. The [TEXTS_PLACEHOLDER] will be replaced by the given statements or atomic claims.

Your task is to identify the logical relationship between the claims in the input statement.
Let's define a function named detect_logic_type(input:str).
The returned value should be a string, where the string is a logic type label that best describes the relationship between the claims in the input.
Use the following predefined logical relationship types: ["causal", "conditional", "concessive", "temporal", "no_logical_relationship"].

For example, if a user calls detect_logic_type("Mary is very tired, so she decided to go to bed early.").
You should return "causal".
Return the logical relationship type without any other words.

Note that your response will be passed to the python interpreter, SO NO OTHER WORDS!

detect_logic_type([STATEMENT_PLACEHOLDER])

Table 21: Instruction used for logical relationship detection. The [STATEMENT_PLACEHOLDER] will be replaced by a given statement.

You are given a statement, a logical relationship type, and an evidence text, and you need to decide whether the evidence supports, refutes, or is irrelevant to the logical relationship claimed in the statement.
Choose from the following three options.
A. The evidence supports the logical relationship.
B. The evidence refutes the logical relationship.
C. The evidence is irrelevant to the logical relationship.

For example, you are give Statement: "Because Mars has a smaller mass than Earth, it rotates slower than Earth.", Logical relationship: "causal", Evidence: "A planet's rotational speed is not directly determined by its mass, but rather by factors such as its initial angular momentum during formation.".
You should return "B".
Pick the correct option "A", "B", "C" without other words.

Statement: [STATEMENT_PLACEHOLDER]
Logical relationship: [LOGICAL_RELATIONSHIP_PLACEHOLDER]
Evidence: [EVIDENCE_PLACEHOLDER]

Table 22: Instruction used for atomic claim verification. The placeholders [STATEMENT_PLACEHOLDER], [LOGICAL_RELATIONSHIP_PLACEHOLDER], and [EVIDENCE_PLACEHOLDER] will be replaced by the given statement, logical relationship type, and evidence.

Your task is to decompose the statement into atomic claims.
Let's define a function named decompose(input:str).
The returned value should be a list of strings, where each string should be a context-independent claim, representing one fact.
For example, if a user call decompose("Mary is a five-year old girl, she likes playing piano and she doesn't like cookies.").
You should return a python list without any other words:
["Mary is a five-year old girl.", "Mary likes playing piano.", "Mary doesn't like cookies."]
Note that your response will be passed to the python interpreter, SO NO OTHER WORDS!

decompose([STATEMENT_PLACEHOLDER])

Table 23: Instruction used for atomic claim decomposition and decontextualization. The [STATEMENT_PLACEHOLDER] will be replaced by a given statement.

You are given a claim and an evidence text, and you need to decide whether the evidence supports, refutes, or is irrelevant to the claim.
Choose from the following three options.
A. The evidence supports the claim.
B. The evidence refutes the claim.
C. The evidence is irrelevant to the claim.

For example, you are give Claim: "Preslav is a professor.", Evidence: "Preslav Nakov is a Professor in MBZUAI NLP group, and also the department chair.".
You should return "A".
Pick the correct option "A", "B", "C" without other words.

Claim: [ATOMIC_CLAIM_PLACEHOLDER]
Evidence: [EVIDENCE_PLACEHOLDER]

Table 24: Instruction used for atomic claim verification. The placeholders [ATOMIC_CLAIM_PLACEHOLDER] and [EVIDENCE_PLACEHOLDER] will be replaced by the given atomic claim and evidence.