Mirroring Human Mindset: Cognitive-Inspired Multi-hop Fact Verification for Large Language Models

Anonymous ACL submission

Abstract

Multi-hop fact verification evaluates the truthfulness of claims by reasoning over multiple facts, serving as a fundamental task in identifying misinformation. Previous approaches tend to directly feed claims to be verified into Large Language Models (LLMs) and reason for an answer, without explicitly teaching LLMs how to comprehend and reason over multiple pieces of evidence. Consequently, these often yield flawed reasoning trajectories and overlook critical evidentiary details. To bridge this gap, we propose **CogFact**, a framework that inspires LLMs by mirroring human mindset for multi-hop facts. Specifically, CogFact consists of three stages that mirror key aspects of human cognitive processes. Firstly, CogFact performs implicit entity resolution over multi-hop claims, then conducts semantic decomposition on each component of the claim, and ultimately constructs an integrative logical chain to reason. CogFact can significantly enhance the multi-hop reasoning capabilities of LLMs while also offering excellent interpretability and scalability. We evaluate CogFact on two public multi-hop fact verification datasets, where it achieves state-of-the-art performance compared to multiple baselines.

1 Introduction

The pervasive dissemination of misinformation on social media has driven an urgent need for fact verification (Park et al., 2021; Botnevik et al., 2020; Si et al., 2023; Liu et al., 2023; Wang et al., 2024). Multi-hop fact verification is a sophisticated fact verification task that entails assessing the veracity of a given claim through multi-step reasoning based on multiple retrieved pieces of evidence (Thorne and Vlachos, 2018; Ma et al., 2023; Zhu et al., 2023; Pan et al., 2023; Aly and Vlachos, 2022; Thorne et al., 2018). Thus, such complex fact verification has attracted extensive attention from researchers due to its closer alignment with real-world misinformation detection scenarios.

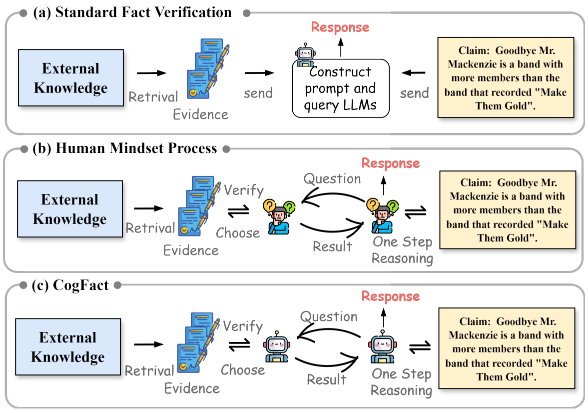


Figure 1: Comparison of CogFact with Standard Fact Verification and Human Mindset Process for Multi-hop Fact Verification. (a) shows the standard fact verification process, where the evidence and claim are jointly fed into a LLMs. (b) illustrates the human mindset process involving reasoning, choosing, and verifying. (c) demonstrates CogFact, which plans the reasoning path for LLMs via Human Mindset.

Concurrently, the technological maturation of Large Language Models (LLMs) has catalyzed significant advancements in fact verification (Pan et al., 2023; Zhang and Gao, 2023; Yue et al., 2023; Si et al., 2024). Early methodologies for multi-hop fact verification often employed reasoning graphs or fine-tuned natural language inference (NLI) models to verify the multi-hop facts (Ren and Leskovec, 2020; He et al.; Lei et al., 2025). In contrast, recent approaches emphasize understanding and reasoning by leveraging inference-time scaling, such as Chain-of-Thought (CoT) and code-based reasoning, to perform multi-hop fact verification (Pan et al., 2023; Zhang and Gao, 2023; Cao et al., 2025; Yao et al., 2024; Wang et al., 2024). However, for complex reasoning tasks, these methods merely instruct the model what to do rather than how to do it, which often results in flawed reasoning paths and the omission of critical details.

As illustrated in Figure 1(a), standard fact verification directly prompts the LLM to produce a verification result. Such methods merely instruct LLMs on what to do without specifying how to execute the verification process. Consequently, LLMs are forced to rely on the own empirical knowledge, which renders the resulting reasoning paths highly unpredictable and uncontrollable. To address this issue, we propose a cognitive-inspired reasoning paradigm to guide the reasoning. As shown in Figure 1(b), human cognition follows an iterative loop that alternates between one step reasoning and verification of the reasoning outcomes. Building on this insight, we develop CogFact, as depicted in Figure 1(c). This framework yields more transparent and controllable reasoning trajectories compared to standard fact verification methods.

Inspired by the human cognitive process, we propose **CogFact**, a novel framework for multi-hop fact verification. CogFact mirrors the human mindset in conducting multi-hop reasoning and fact-checking processes, effectively steering LLMs along factually correct reasoning paths. Specifically, CogFact is divided into three stages. The first stage, *Implicit Entity Resolution*, involves disambiguating entity descriptions and mapping to the original entity. By simulating human-like entity recognition, CogFact enables LLMs to accurately identify and reason about implicit entities within complex claims. In the second stage, *Semantic Decomposition*, the framework instructs LLMs to decompose the claim into a sequence of atomic questions. These questions serve as targeted queries to an external knowledge base, ensuring that the LLM scrutinizes fine-grained details by addressing each semantic segment individually. Finally, in the *Integrative Logical Reasoning* stage, CogFact integrates the above outputs to derive the final conclusion through reasoning logic that aligns with human cognition. Compared to previous methods, CogFact has more accurate reasoning paths, pays more attention to details, and is less prone to hallucination.

We evaluate CogFact on two challenging multi-hop fact verification datasets: HOVER (Jiang et al., 2020) and EX-FEVER (Ma et al., 2023). Experimental results demonstrate that CogFact consistently outperforms nine competitive baselines, achieving state-of-the-art performance on both benchmarks. Furthermore, as discussed in Section 4.3 and Section 4.4, CogFact exhibits robust scalability and interpretability. In summary, our primary contributions are as follows:

1. We review current fact verification methods and observe that the exclusive reliance on LLMs intrinsic verification ability increases the likelihood of flawed reasoning trajectories.

2. We propose CogFact, a framework that provides procedural guidance to LLMs by mirroring the human mindset, thereby ensuring a structured and controllable reasoning path.

3. Our experimental results demonstrate that CogFact achieves superior performance compared to existing methodologies.

2 Related Work

In recent years, the automation of fact verification has garnered widespread attention (Park et al., 2021; Botnevik et al., 2020; Si et al., 2023; Liu et al., 2023; Cekinel et al., 2025), and an increasing number of researchers are delving into the identification of misinformation (Lee et al., 2021; Mehta et al., 2022; Zhou et al., 2023; Liu et al., 2025). Many studies have focused on improving factual understanding and reasoning capabilities. (Hanselowski et al., 2018) implement an entity-linking and use the Enhanced LM for ranking candidate facts and classification. (Zhang and Gao, 2023) utilize GPT to partition the claim and unearth the underlying key points within the claim using external knowledge.

Moreover, some researchers argue that single-hop reasoning is insufficient, as real-world claims may encompass information from multiple sources (Jiang et al., 2020; Pan et al., 2023; Yang et al., 2018). Therefore, the task of multi-hop fact verification has been proposed (Jiang et al., 2020; Ma et al., 2023; Aly and Vlachos, 2022; Pan et al., 2023). For example, (Zhu et al., 2023) utilize a counterfactual approach to obtain additional multi-hop data, enhancing the generalization. ProgramFC (Pan et al., 2023) relies on LLMs to generate static program templates and perform reasoning based on these templates. Meanwhile, (Liu et al., 2025) integrates multiple role-playing LLMs to verify factual claims.

However, the aforementioned methods merely require LLMs to solve the problem or its subproblems, without explicitly teaching them how to reason. In contrast, our CogFact mitigates the risk of erroneous reasoning paths by mirroring human thinking patterns and endowing LLMs with cognitively grounded reasoning trajectories.

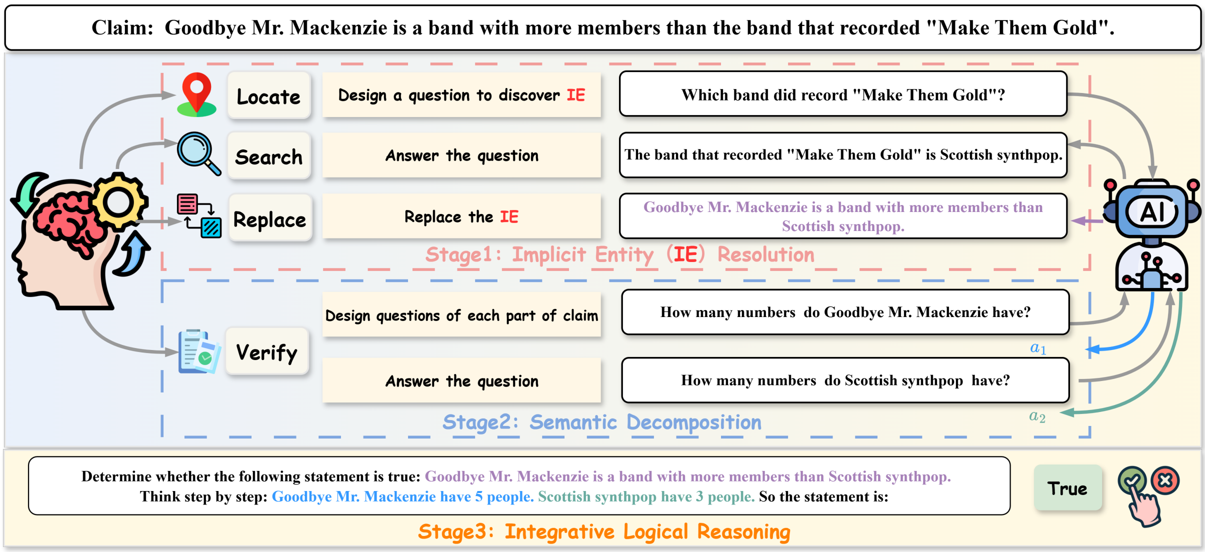


Figure 2: Overview of CogFact. CogFact is divided into three stages: Implicit Entity Resolution, Semantic Decomposition, and Integrative Logical Reasoning. The reasoning path follows a top-down progression. The purple, blue, and green segments in Integrative Logical Reasoning correspond to the outputs of the first two stages.

3 Methodology

3.1 Preliminary

The multi-hop fact verification can be formulated as follows: given a claim c, which is associated with numerous pieces of evidence, the set $E = \{e_1, e_2, ...e_n\}$ denotes the relevant evidence in a large textual corpus such as Wikipedia. The target of multi-hop fact verification is to assess whether a claim is supported by a set of evidence through a certain reasoning process. If supported, the claim is deemed true; otherwise, it is considered false, we use y to denote the final label of the claim, where $y \in \{True, False\}$. Since our task focuses on understanding and reasoning about the retrieved evidence, the evidence E that we utilize is considered gold evidence, i.e., $e_i \in \{e_1, e_2....e_n\}$ that can either support or refute the claim.

3.2 CogFact

Figure 2 illustrates the overall architecture of the CogFact framework, which is organized into three stages: Implicit Entity Resolution, Semantic Decomposition, and Integrative Logical Reasoning.

3.2.1 Implicit Entity Resolution

In human cognition, the reasoning process begins by identifying the implicit entities within a claim. We abstract this cognitive process as follows. Let $T = \{t_1, t_2, t_3, \ldots, t_m\}$ denote the components of a claim associated with external evidence, where t_i represents a word in claim and T typically represents a description of an entity. For example, in the claim "The Rookie of The Year in the 1997 CART season drives it in the NASCAR Sprint Cup Series," $T = \{$"The", "Rookie", "of", "The", "Year", "in", "the", "1997", "CART", "season"$\}$. Implicit Entity Resolution involves replacing the entity description T with its represented entity T'. To implement this, we propose three components $(\mathcal{L}, \mathcal{S}, \mathcal{R})$, representing locate, search, and replace operations.

Locate ($\mathcal{L}$): This component identifies the span of the related part T within a claim c. To facilitate downstream reasoning, the system guides the LLM to generate a targeted question $Q_l = \mathcal{L}(c)$ regarding the entity description. For instance, in Figure 2, for the related part "the band that recorded 'Make Them Gold'", LLMs would generate the question: "Which band did record 'Make Them Gold'?" Specifically, we generate prompts in the format demonstrated as follows: "<claim> Design a question to discover the implicit entity. If no implicit entity is found, print: 'No implicit entity'". We utilize the same format with k examples as a few-shot in the prompt. Due to different types of graph shapes of multi-hop reasoning (Jiang et al., 2020), to find the optimal reasoning path for identifying implicit entities, we employ dynamic sample selection (Nori et al., 2023). As illustrated in Figure 3, the claim c is mapped to an embedding e_c, which is then used to retrieve the top k most semantically similar instances from a curated train-

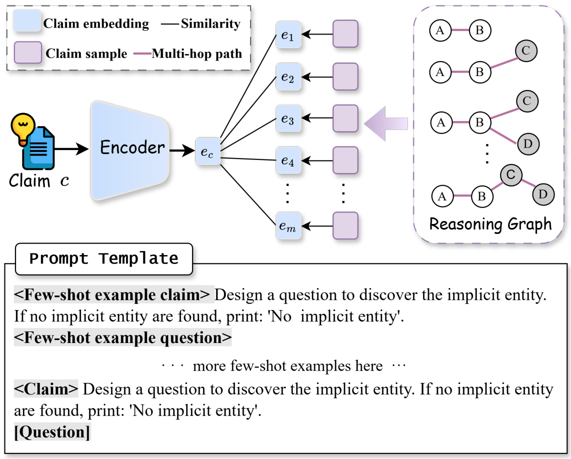


Figure 3: Dynamic sample selection approach and prompt template. Where [Question] denotes that LLMs need to generate questions for the relevant part.

ing library $\{e_1, \ldots, e_m\}$. These retrieved instances are subsequently integrated into context-aware few-shot prompts designed to optimize the localization process.

Search ($\mathcal{S}$)**:** Given the generated query Q_l, this component accesses external evidence to derive an answer A_l that explicitly contains the identified entity T'. This operation is denoted as $A_l = \mathcal{S}(Q_l, \text{Evidence})$. Specifically, we generate prompts in the following format: "<evidence> Based on the above facts, answer the question. <question>". Subsequently, LLMs will return the answer A_l to the question Q_l, and A_l includes the identified implicit entity T'.

Replace ($\mathcal{R}$)**:** This final operation leverages LLMs to replace the original description T in claim c with the entity T' found in A_l. The constructed prompt format is as follows: "<Answer> on the above information, replace the implicit entity and its description in the following expressions with specific names: <claim>". To facilitate the learning of the replacement process by LLMs, we also adopt dynamic sample selection. Ultimately, we obtain a claim $c' = \mathcal{R}(c, A_l)$ that has undergone the replacement process.

3.2.2 Semantic Decomposition

After implicit entity resolution, humans tend to verify facts by carefully examining each individual detail. Accordingly, we introduce semantic decomposition to prevent LLMs from overlooking critical details. Like human mindset, semantic decomposition is to pose questions Q_s for each part of claim c' and obtain relevant answers from external knowledge, where $Q_s = \{q_1, q_2, q_3, ...q_n\}$, $q_i, i \in [1, n]$ represents a question. For the claim "Greater Swiss Mountain Dog and Harrier are both dog breeds." questions can be posed for each complete semantic segment, such as "Is it true that Greater Swiss Mountain Dog are dog breeds?" and "Is it true that Harrier are dog breeds?" In Figure 2, we also generate two questions for a claim.

Subsequently, we verify the relevant answers for each question against the evidence repository. Similar to the approach in the first stage, we inject evidence into LLMs to provide external knowledge. We then pose direct question q_i to LLMs to obtain the answer $a_i, i \in [1, n]$ respectively. All the answers are represented as the set $A_s = \{a_1, a_2, ...a_n\}$. The specific construction of prompts is as follows: "<evidence> Base on the above facts, <question>".

3.2.3 Integrative Logical Reasoning

Ultimately, for reasoning that follows human cognitive logic, we manually construct chain-of-thought templates and integrating the outputs from the aforementioned two stages as part of these templates. Then, LLMs can then review the prior reasoning process and reason the final answer solely based on the provided content. Specifically, the chain-of-thought (CoT) template base on c' and A_s and guide the reasoning process in a human-like manner. The prompt for constructing the CoT is: "Determine whether the following statement is true: <c'> True or false? Think step by step: <A_s> So, the statement is:". In the end, LLMs provide the answer $y \in \{True, False\}$. In this stage, LLMs focus on reasoning without requiring external knowledge.

4 Experiments

Datasets. To better align with the requirements of real-world fact verification, our experiments are conducted on two multi-hop fact verification datasets, HOVER (Jiang et al., 2020) and EX-FEVER (Ma et al., 2023). HOVER is a multi-hop dataset derived from English Wikipedia articles. We utilize the validation set of HOVER for evaluation, which consists of 4,000 claims. The claims in HOVER require evidence from up to four Wikipedia articles to determine whether the claims are true. EX-FEVER involves 2-hop and 3-hop reasoning, claims are created by summarizing and modifying information from hyperlinked Wikipedia documents. We evaluate the model using

Table 1: Main results (%). Macro-F1 scores of CogFact and baselines in the few-shot setting. Results in bold are the best performance.

Few-shot Model	HOVER				EX-FEVER		
	2-hop	3-hop	4-hop	Total	2-hop	3-hop	Total
I. Vanilla pretrained model							
BERT (Kotonya and Toni, 2020)	52.32	53.02	50.53	51.94	64.41	57.89	61.15
FLAN-T5 (Chung et al., 2022)	64.23	56.39	54.19	56.41	72.42	61.72	67.64
II. Inference augmented model							
Scandi-NLI (Alexandrainst, 2022)	72.98	64.85	57.66	65.35	65.66	59.48	62.75
DeBERTaV3-NLI (Laurer et al., 2024)	73.93	63.88	51.69	64.20	69.25	61.83	66.10
ProgramFC (Pan et al., 2023)	74.99	66.11	63.97	68.05	82.87	77.93	80.54
III. LLM-based model							
GPT-3.5 (Brown et al., 2020)	73.27	61.79	54.90	63.23	81.01	72.31	76.71
GPT-3.5-CoT (Wei et al., 2022)	71.58	64.47	62.56	65.97	79.37	74.23	76.65
HiSS (Zhang and Gao, 2023)	73.45	65.10	63.91	67.14	82.28	74.91	78.61
Factcheck-GPT (Wang et al., 2024)	75.15	66.01	64.33	68.69	82.81	75.10	78.96
CogFact	**76.49**	**69.43**	**65.82**	**70.47**	**84.37**	**78.92**	**81.50**

the test set of EX-FEVER. To maintain consistency with HOVER, we remove data with the NEI label. After that, the number of claims is 4,071.

Baselines. We use nine baselines for comparison with CogFact, which can be categorized into three categories. (I). Vanilla pretrained model, BERT (Kotonya and Toni, 2020) simultaneously places the claim and evidence into a pre-trained BERT encoder for a classification task. Flan-T5 (Chung et al., 2022) is a unified Text-to-Text Transformer. We implemented classification for the few-shot examples by leveraging the ICL. We create a prompt by combining the claim and evidence and obtain the classification result through the Text-to-Text approach. (II). Inference augmented model, ScandiNLI (Alexandrainst, 2022) fine tunes nb-bert-large for Natural Language Inference (NLI). DeBERTaV3-NLI (Laurer et al., 2024) fine tunes the DeBERTaV3 model on annotations from FEVER and four NLI datasets. ProgramFC (Pan et al., 2023) uses a shared library of specialized functions to reason with the help of codex and Flan T5. (III). LLM-based model, based on the gpt-3.5-turbo engine. The reasoning is based on the few-shot prompting, where using ICL from the training set before the test instance. For GPT-3.5 and GPT-3.5-CoT, we inquire about the reasoning results of LLMs using standard prompts and prompts in the form of CoT, respectively. The HiSS (Lee et al., 2021) uses LLMs to partition the claim and determine the truth of each subpart. Factcheck-GPT (Wang et al., 2024) is a fine-grained system for fact-checking. For fairness, we impose a constraint that the model cannot access the web.

Implementation Details. For all models, we employ the few-shot setting, with $k = 16$ chosen as the number of few-shot examples or fine-tune data. For the fine-tune models, we use cross-entropy loss and the AdamW optimizer with a learning rate of 1e-5. We employ gpt-3.5-turbo-0613 as the backbone for the GPT-3.5-based model. To stabilize the output of model, we use a greedy decoding strategy, setting the temperature to 0. We also use the FLAN-T5-XL 3B as the backbone of the T5 module.

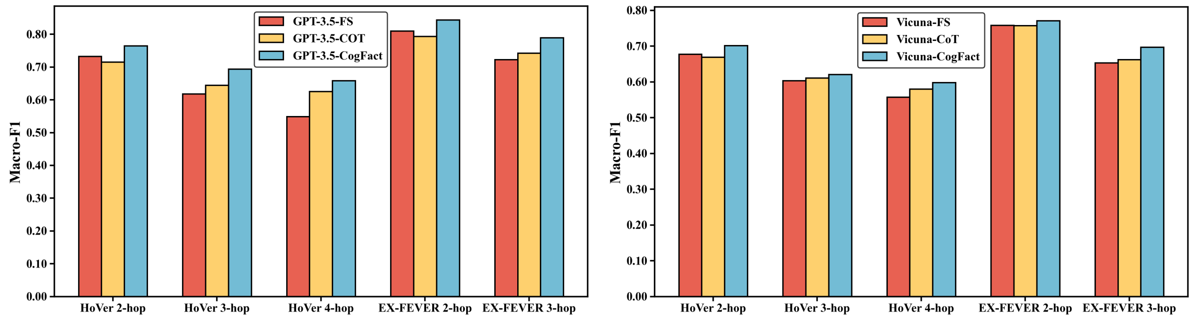
4.1 Main Results and Discussion

As shown in Table 1, we demonstrate the performance of CogFact and baselines. Furthermore, we can observe the following findings:

Compared to directly asking LLMs to perform reasoning, explicitly teaching them how to reason is more beneficial. We find that the performance of GPT-3.5 enhanced with CoT is not remarkable, even less than the standard prompt GPT-3.5 by 1.69% and 1.64% on 2-hop claims. Compared to CogFact, the performance dropped by 4.50% and 4.85% on both datasets. This suggests that the requirement for extensive external knowledge in multi-hop fact verification reasoning makes it challenging for LLMs to find reasonable reasoning paths. Therefore, the current effectiveness of the CoT method in the fact verification domain is not significant. In comparison, CogFact can rapidly aid LLMs reasoning through human mindset, resulting in more rational reasoning paths.

Table 2: Ablation results (%). Macro-F1 scores for CogFact and the ablation models.

	HOVER			EX-FEVER	
	2-hop	3-hop	4-hop	2-hop	3-hop
w\o Dynamic Sample Selection	76.68	68.77	64.93	84.34	77.68
w\o Implicit Entity Resolution	73.09	68.34	64.57	84.01	77.84
w\o Semantic Decomposition	73.06	61.75	60.04	75.02	65.55
CogFact	76.49	69.43	65.82	84.37	78.92



(a) CogFact and baselines based on GPT-3.5

(b) CogFact and baselines based on Vicuna

Figure 4: Macro-F1 scores of different LLMs based CogFact and baselines on two datasets.

Models with enhanced inference capabilities have an advantage when dealing with complex claims. Among the baselines, Scandi-NLI and DeBERTaV3-NLI outperform BERT and Flan-T5 models that have not been pre-trained on natural language inference data on HOVER and EX-FEVER. This performance gap is most evident in the comparison between GPT-3.5-CoT and GPT-3.5. While the former shows a slight 1.69% decrease on 2-hop claims, it achieves significant gains of 2.68% and 7.66% on 3-hop and 4-hop claims, respectively. Furthermore, CogFact, which leverages human-inspired reasoning paradigms, demonstrates even more substantial improvements, outperforming baselines by up to 15.29% and 20.35% on 4-hop claims.

CogFact is more flexible and reasoning rationally. Compared to static program-based ProgramFC (Pan et al., 2023), CogFact shows improvements of 2.42% and 0.96% on both datasets. Similarly utilizing LLMs, this indicates that LLMs guided by human mindset lead to more rational paths for fact verification and are more flexible. HiSS (Zhang and Gao, 2023) and Factcheck-GPT (Wang et al., 2024) only focuses on dividing the claim without emphasizing reasoning to determine the truthfulness. As a result, the performance on both datasets is approximately 2% lower than that of CogFact.

4.2 Ablation Study

In this section, we conduct an ablation analysis of the CogFact, investigating the impact of three modules: Dynamic Sample Selection, Implicit Entity Resolution, and Semantic Decomposition. The results are shown in Table 2.

Effect of Dynamic Sample Selection. Dynamic sample selection is helpful for CogFact. As shown in Table 2, removing dynamic sample selection results in a 0.19% improvement in the performance on 2-hop claims and leading to a decrease in accuracy of 0.66% and 0.89% on the 3-hop and 4-hop claims on HOVER, respectively. Actually, this outcome was anticipated. Dynamic sample selection assists the CogFact in selecting similar claims as few-shot examples, where similar claims and the target claim share similar reasoning structures, thereby making CogFact more adept at finding the correct reasoning path. In simpler claims, such as 2-hop claims, the effect may not be significant. However, in more complex claims, similar structure few-shot examples can greatly aid CogFact in completing the entire solving process, leading to a significant improvement.

Effect of Implicit Entity Resolution. The absence of Implicit Entity Resolution complicates the subsequent Semantic Decomposition stage due to increased task coupling. As evidenced by the ablation study in Table 2, removing this module leads to an average performance decline of 1.91% and 0.72%, underscoring its necessity in the overall pipeline.

Effect of Semantic Decomposition. We directly use the evidence to judge the veracity of the claim after entity reconstruction, and the results are presented in Table 2. In comparison, performance of CogFact was significantly impacted (decrease ranging from 3.43% to 13.17%). Firstly, the loss of the Semantic Decomposition resulted in an inability to perceive details, causing CogFact to be more inclined towards the overall truthfulness of the claim. Secondly, due to the absence of information about each part, we could not construct CoT for reasoning, thus affecting its reasoning ability. Therefore, the performance suffers a significant decrease.

4.3 CogFact Can Help Different LLMs

CogFact is designed as a general framework capable of enhancing various LLMs. To evaluate its generalizability, we conducted experiments using both GPT-3.5 and Vicuna (Chiang et al., 2023). Vicuna is an open-source chatbot trained by fine-tuning LLaMA on user-shared conversations collected from ShareGPT. Specifically, we employed two LLMs to validate the effectiveness of CogFact and we compared CogFact against two primary baselines: Few-shot mode (FS) and Chain-of-Thought (CoT). Figure 4(a) illustrates the performance of the GPT-3.5-based model, while Figure 4(b) presents the performance of the Vicuna based model. The results demonstrate that LLMs enhanced with the CogFact exhibit higher precision compared to the few-shot and CoT methods. Moreover, the relative decrease in performance of CogFact is less pronounced as the number of hops increases. These findings confirm that CogFact effectively bolsters the multi-hop fact verification reasoning capabilities of various LLMs.

4.4 The Interpretative Role of Integrative Logical Inference

EX-FEVER dataset provides golden explanations for the label, which is a textual explanation that describes the minimally sufficient information in each hop to verify a claim. To verify the interpretability of the CogFact, we conducted experiments using the golden explanations data from the EX-FEVER dataset. We consider that the constructed CoT, as an intermediate result of CogFact, is accessible and can provide explanations for the judgments. We use the rouge metrics to measure the matching degree between the constructed CoT and the golden explanation. Ultimately, we found that Constructed CoT has an explanatory role. The experimental results are shown in Table 3. We compared CogFact with three models: MDR (Xiong et al., 2021), BERT (Devlin et al., 2018), and GPT (Brown et al., 2020), as used in the study by (Ma et al., 2023). Although our model is not explicitly designed for generating explanations, it outperforms three models on average by 3.8% and 9.19% in Rouge-1 and Rouge-L metrics. Additionally, we demonstrate the interpretability of CogFact through an example, as shown in Table 4, where CogFact accurately identifies claims that do not align with the facts. More cases of interpretability can be found in Appendix A.1.

Table 3: Results of different models generating explanations (%). Results in bold are the best performance.

	MDR	BERT-based	GPT	CogFact
Rouge-1	54.88	46.88	52.28	**55.15**
Rouge-2	**41.34**	32.80	33.74	35.38
Rouge-L	49.42	35.52	48.13	**53.55**

4.5 Error Type Analysis

To better understand the behavior of our CogFact and facilitate future research, we randomly sample 40 examples that are generated by CogFact. As shown in Figure 5(a), we categorize the causes of errors into three types: 1) *logical mistake*, 2) *factual hallucination*, and 3) *thoughts omission*. Cases corresponding to each type can be found in Appendix A.2.

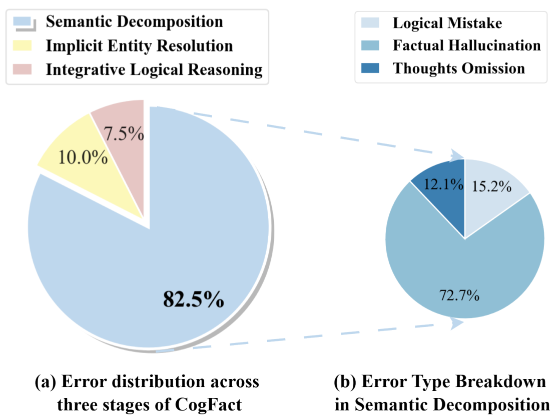


Figure 5: Error analysis of CogFact on 40 samples.

We find that the errors corresponding to the three stages are distributed as 10%, 82.5%, and 7.5%, respectively. Consequently, we further analyzed the

Table 4: Examples of CogFact solving claims on HOVER. The blue text indicates the flawed part of the claim.

Claim: William Boyd was British and was born on June 23, 1967. He and Jim Thompson were both screenwriters.
<No implicit entities were found.>
q_1: Is it true that William Boyd was British? q_2: Is it true that William Boyd was born on June 23, 1967? q_3: Is it true that William Boyd and Jim Thompson were both screenwriters?
a_1: Yes, it is true that William Boyd is a British writer. a_2: No, it is not true that William Boyd was born on June 23, 1967. William Boyd was born on March 7, 1952. a_3: Yes. it is true that both William Boyd and Jim Thompson were screenwriters in addition to being authors.
LLM input: Determine whether the following statement is true: William Boyd was British and was born on June 23, 1967. He and Jim Thompson were both screen-writers. Think step by step: it is true that William Boyd is a British writer. It is not true that William Boyd was born on June 23, 1967. William Boyd was born on March 7, 1952. it is true that both William Boyd and Jim Thompson were screenwriters in addition to being authors. So the statement is: **LLM output:** False **Ground Truth:** False

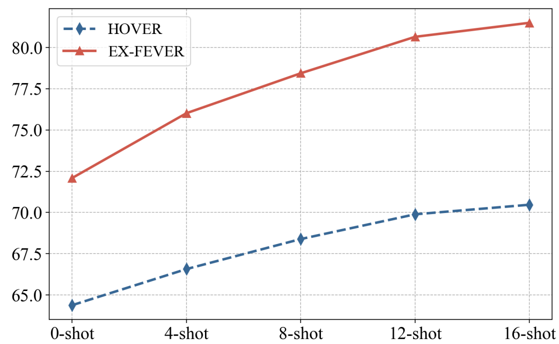


Figure 6: Results (%). Effect of few-shot number.

types of errors occurring in the Semantic Decomposition stage. As shown in Figure 5(b), 72.7% of the errors are concentrated in hallucinations generated by LLM itself. Meanwhile, although decomposing sub-information enhances the logical reasoning of LLMs, it also ignores longer dependencies in the text, resulting in some degree of performance loss. In the CoT construction phase, there is still room for improvement in the reasoning logic of LLMs.

4.6 The Effect of Few-shot Quantity

We conduct experiments to demonstrate the impact of the number of few-shot samples used on the CogFact. In Figure 6, it can be observed that with the increase in the number of few-shot instances, the performance of CogFact also increases but tends to stabilize. Due to the length limitation of GPT-3.5 input, we ultimately utilize 16 samples as few-shot instances. This phenomenon aligns with the general trend observed in the use of LLMs through ICL. When the number of few-shot cases is low, there may be issues such as irregular response formats leading to performance degradation. However, as the number gradually increases, the performance gradually reaches normal levels.

5 Conclusion

In this paper, we identify a common limitation in existing fact-checking methods, which merely require LLMs to perform reasoning, without explicitly instructing LLMs on how to reason. Therefore, we propose CogFact, a novel framework for multi-hop fact verification. CogFact effectively steers LLMs along logically sound and factually correct reasoning paths by mirroring the human mindset. Then, our experiments demonstrate that the correctness of reasoning paths leads to significant improvements in model performance and that CogFact achieves state-of-the-art performance on both datasets. Moreover, CogFact exhibits good interpretability and scalability, as the reasoning process is explicitly visible, and empirical results validates its efficacy in augmenting the capabilities of various LLMs.

6 Limitation

Despite its superior performance, CogFact has the limitation that warrant further investigation. We think that its computational intensity is relatively high. This latency is inherently linked to the complexity of the human-like reasoning process, which necessitates multiple steps to ensure a rigorous conclusion. Currently, our approach requires approximately 5 to 6 times more processing time than direct LLM querying. Future work will focus on optimizing the inference efficiency to achieve a better balance between accuracy and latency.

7 Ethical Considerations

Our methodology acknowledges the potential presence of biases in the data used to train LLMs, including in the assessment of factuality. These biases are inherent and beyond our direct control.

Our models are intended for broad public engagement and have the potential to significantly assist human fact-checkers by streamlining their workflows. However, we recognize the risk that these tools could be exploited for malicious purposes. We urge the researchers to approach the use of these models with caution and integrity.

The computational demands of training and running LLMs contribute to energy consumption, with implications for global warming. Our methodology, which relies on few-shot learning, lessens the environmental impact by not requiring the training from the ground up. Nevertheless, it is important to note that the GPT-3.5 model, accessed via API for inference, still has a substantial energy footprint.

References

Alexandrainst. 2022. Scandinli: Natural language inference for the scandinavian languages. https://github.com/alexandrainst/ScandiNLI. Accessed: 2024-03-16.

Rami Aly and Andreas Vlachos. 2022. Natural logic-guided autoregressive multi-hop document retrieval for fact verification. *arXiv preprint arXiv:2212.05276*.

Bjarte Botnevik, Eirik Sakariassen, and Vinay Setty. 2020. Brenda: Browser extension for fake news detection. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 2117–2120.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Han Cao, Lingwei Wei, Wei Zhou, and Songlin Hu. 2025. Enhancing multi-hop fact verification with structured knowledge-augmented large language models. AAAI'25/IAAI'25/EAAI'25. AAAI Press.

Recep Firat Cekinel, Pinar Karagoz, and Çağrı Çöltekin. 2025. Multimodal fact-checking with vision language models: A probing classifier based solution with embedding strategies. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4622–4633, Abu Dhabi, UAE. Association for Computational Linguistics.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and 1 others. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. UKP-athene: Multi-sentence textual entailment for claim verification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 103–108, Brussels, Belgium. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. Hover: A dataset for many-hop fact extraction and claim verification. *arXiv preprint arXiv:2011.03088*.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. *arXiv preprint arXiv:2010.09926*.

Moritz Laurer, Wouter Van Atteveldt, Andreu Casas, and Kasper Welbers. 2024. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. *Political Analysis*, 32(1):84–100.

Nayeon Lee, Yejin Bang, Andrea Madotto, Madian Khabsa, and Pascale Fung. 2021. Towards few-shot fact-checking via perplexity. *arXiv preprint arXiv:2103.09535*.

Deren Lei, Yaxi Li, Siyao Li, Mengya Hu, Rui Xu, Ken Archer, Mingyu Wang, Emily Ching, and Alex Deng. 2025. Factcg: Enhancing fact checkers with graph-based multi-hop data.

Fei Liu, Yibo Li, and Meiyun Zuo. 2023. Keshem: Knowledge enabled short health misinformation detection framework. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 372–388. Springer.

Yuxuan Liu, Hongda Sun, Wenya Guo, Xinyan Xiao, Cunli Mao, Zhengtao Yu, and Rui Yan. 2025. Bidev: bilateral defusing verification for complex claim fact-checking. AAAI'25/IAAI'25/EAAI'25. AAAI Press.

Huanhuan Ma, Weizhi Xu, Yifan Wei, Liuji Chen, Liang Wang, Qiang Liu, and Shu Wu. 2023. Ex-fever: A dataset for multi-hop explainable fact verification. *arXiv preprint arXiv:2310.09754*.

Nikhil Mehta, María Leonor Pacheco, and Dan Goldwasser. 2022. Tackling fake news detection by continually improving social context representations using graph neural networks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1363–1380.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, and 1 others. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. *arXiv preprint arXiv:2305.12744*.

Jungsoo Park, Sewon Min, Jaewoo Kang, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2021. Faviq: Fact verification from information-seeking questions. *arXiv preprint arXiv:2107.02153*.

Hongyu Ren and Jure Leskovec. 2020. Beta embeddings for multi-hop logical reasoning in knowledge graphs. In *Neural Information Processing Systems*.

Jiasheng Si, Yibo Zhao, Yingjie Zhu, Haiyang Zhu, Wenpeng Lu, and Deyu Zhou. 2024. CHECKWHY: Causal fact verification via argument structure. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15636–15659, Bangkok, Thailand. Association for Computational Linguistics.

Jiasheng Si, Yingjie Zhu, and Deyu Zhou. 2023. Exploring faithful rationale for multi-hop fact verification via salience-aware graph learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13573–13581.

James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. *arXiv preprint arXiv:1806.07687*.

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.

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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Wenhan Xiong, Xiang Li, Srini Iyer, Jingfei Du, Patrick Lewis, William Yang Wang, Yashar Mehdad, Scott Yih, Sebastian Riedel, Douwe Kiela, and Barlas Oguz. 2021. Answering complex open-domain questions with multi-hop dense retrieval. In *International Conference on Learning Representations*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Zhenrui Yue, Huimin Zeng, Yang Zhang, Lanyu Shang, and Dong Wang. 2023. Metaadapt: Domain adaptive few-shot misinformation detection via meta learning. *arXiv preprint arXiv:2305.12692*.

Xuan Zhang and Wei Gao. 2023. Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method. *arXiv preprint arXiv:2310.00305*.

Yangming Zhou, Yuzhou Yang, Qichao Ying, Zhenxing Qian, and Xinpeng Zhang. 2023. Multimodal fake news detection via clip-guided learning. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2825–2830. IEEE.

Yingjie Zhu, Jiasheng Si, Yibo Zhao, Haiyang Zhu, Deyu Zhou, and Yulan He. 2023. Explain, edit, generate: Rationale-sensitive counterfactual data augmentation for multi-hop fact verification. *arXiv preprint arXiv:2310.14508*.

A Appendix

A.1 Correct Examples of Case Studies

We have presented correct examples that were randomly sampled to aid in comprehending the behavior of CogFact. These examples demonstrate how the model effectively identifies implicit entities and performs multi-step verification to ensure factual consistency. By decomposing a complex claim into atomic sub-questions, CogFact reduces the risk of overlooking hidden factual errors that are often missed by vanilla CoT prompting.

In the examples of case studies, q_l, a_l, and c correspond to the components of the *Implicit Entity Resolution* stage; Q_i and its corresponding response belong to the *Semantic Decomposition* stage; and the final LLM input and output correspond to the *Integrative Logical Reasoning* stage.

Correct Case Study

Claim: Bruno Tonioli is an Italian choreographer. He appeared on Season 12 of the television show that featured Ralph Macchio.

q_l: Which television show featured Ralph Macchio on Season 12?

a_l: The television show that featured Ralph Macchio on Season 12 is "Dancing with the Stars."

c: Bruno Tonioli is an Italian choreographer. He appeared on Season 12 of "Dancing with the Stars," the television show that featured Ralph Macchio.

- **Q1:** Is it true that Bruno Tonioli is an Italian choreographer? → **Yes, it is true that Bruno Tonioli is an Italian choreographer.**
- **Q2:** Is it true that Bruno Tonioli appeared on Season 12 of "Dancing with the Stars"? → **Yes, it is true that Bruno Tonioli appeared as a judge on Season 12 of "Dancing with the Stars".**
- **Q3:** Is it true that "Dancing with the Stars" featured Ralph Macchio? → **Yes, it is true that Ralph Macchio competed on the twelfth season of "Dancing with the Stars."**

LLM Input: *Constructed CoT* **LLM Output:** **True**
Ground Truth: True

Correct Case Study

Claim: Belgian-born Danny Dare did the choreography for The New Irving Berlin Musical.

◇ *No implicit entities were found.*

- **Q1:** Is it true that Danny Dare is Belgian-born? → **No, it is not true that Danny Dare is Belgian-born. There is no information available to suggest that Danny Dare, the choreographer of the 1942 film "Holiday Inn," is Belgian-born.**
- **Q2:** Is it true that Danny Dare did the choreography for The New Irving Berlin Musical? → **No, it is not true that Danny Dare did the choreography for The New Irving Berlin Musical. The choreography for the musical was not specifically mentioned in the provided information.**

LLM Input: *Constructed CoT* **LLM Output:** **False**
Ground Truth: False

A.2 Incorrect Examples of Case Studies

To better understand the behavior of the CogFact, we randomly selected 40 incorrect samples for manual investigation. We categorized them into three types: 1) *logical mistake*, 2) *factual hallucination*, and 3) *thoughts omission*.

The majority of errors were hallucinations, where the model incorrectly associates Arthur Noss with WWI. Although CogFact and CoT have enhanced reasoning capabilities, logical mistakes still occur. For instance, despite all responses to the questions being 'yes', the model still considers the claim to be false. Also, the elimination of entity questions can also lead the model to forget some important relationships, as shown in the examples.

Incorrect Case Study: Logical Mistake

Claim: Both the Battle of Saipan and the Battle of Manila were fought by American forces.

◇ *No implicit entities were found.*

- **Q1:** Was the Battle of Saipan fought by American forces? → **Yes, the Battle of Saipan was fought by American forces. The U.S. 2nd Marine Division, 4th Marine Division, and the Army's**

27th Infantry Division were involved in the battle and successfully defeated the 43rd Infantry Division of the Imperial Japanese Army.

- **Q2:** Was the Battle of Manila fought by American forces? → **Yes, the Battle of Manila was fought by American and Filipino forces against Japanese troops.**

LLM Input: *Constructed CoT* **LLM Output: False**
Ground Truth: True

Incorrect Case Study: Factual Hallucination

Claim: Arthur Noss was a gunner for the commader during the two European theatre battles Battle of Britain and the Battle of Malta in WWII.

q_l: Who was the commander during the Battle of Britain and the Battle of Malta in WWII?

a_l: The commander during the Battle of Britain and the Battle of Malta in World War II was Air Chief Marshal Sir Keith Rodney Park.

c: Arthur Noss was a gunner for Air Chief Marshal Sir Keith Rodney Park during the Battle of Britain and the Battle of Malta in World War II.

<The elimination progress have been omitted due to space constraints.>

- **Q1:** Is it true that Arthur Noss was a gunner during the Battle of Britain? → No, it is not true that Arthur Noss was a gunner during the Battle of Britain. Arthur Noss served as Keith Park's gunner during World War I, not during the Battle of Britain in World War II.

- **Q2:** Is it true that Arthur Noss was a gunner during the Battle of Malta in World War II? → No, it is not true that Arthur Noss was a gunner during the Battle of Malta in World War II. Arthur Noss was a British World War I flying ace who served as Keith Park's gunner during that time. The Battle of Malta took place during World War II, and Arthur Noss had already been killed in a flying accident during World War I.

- **Q3:** Was Air Chief Marshal Sir Keith Rodney Park involved in the Battle of Britain? → **Yes, Air Chief Marshal Sir Keith Rodney Park was indeed involved in the Battle of Britain. He played a crucial role as the commander of No. 11 Group RAF, which was responsible for the defense of London and southeast England during the battle. His leadership and strategic decisions were instrumental in the successful defense against the German Luftwaffe during this critical phase of the Second World War.**

- **Q4:** Was Air Chief Marshal Sir Keith Rodney Park involved in the Battle of Malta in World War II? → **Yes, Air Chief Marshal Sir Keith Rodney Park was indeed involved in the Battle of Malta during World War II. He played a crucial role in the defense of Malta against Axis forces, helping to secure victory in this significant air battle.**

LLM Input: *Constructed CoT* **LLM Output: False**
Ground Truth: True

Incorrect Case Study: Thoughts Omission

Claim: Miles Alexander Teller is an American actor who appeared in a 2013 coming-of-age romantic drama film featuring an American actress and filmmaker in leading roles.

q_l: Who is the American actress and filmmaker in leading roles in the 2013 coming-of-age romantic drama film?

a_l: The American actress in a leading role in the 2013 coming-of-age romantic drama film "The Spectacular Now" is Shailene Woodley, not Brie Larson.

c: Miles Alexander Teller is an American actor who appeared in "The Spectacular Now" (2013), a coming-of-age romantic drama film featuring Shailene Woodley in a leading role.

<The information about filmmaker was omitted.>

<The following questions and answers have been omitted due to space constraints.>

- **Q1:** Is it true that Miles Alexander Teller is an American actor? $\rightarrow$ **Yes, it is true that Miles Alexander Teller is an American actor.**
- **Q2:** Is it true that Miles Alexander Teller appeared in "The Spectacular Now" (2013)? $\rightarrow$ **Yes, it is true that Miles Teller appeared in "The Spectacular Now" (2013).**
- **Q3:** Is "The Spectacular Now" (2013) a coming-of-age romantic drama film? $\rightarrow$ **Yes, "The Spectacular Now" (2013) is a coming-of-age romantic drama film. It follows the story of two high-schoolers, Sutter and Aimee, as they navigate unexpected encounters and a blossoming romance. The film received critical acclaim and is known for its portrayal of young love and personal growth.**
- **Q4:** Does "The Spectacular Now" (2013) feature Shailene Woodley in a leading role? $\rightarrow$ **Yes, "The Spectacular Now" (2013) features Shailene Woodley in a leading role. She plays the character Aimee, one of the two main high-schoolers in the film.**

LLM Input: *Constructed CoT* **LLM Output: True**
Ground Truth: False