A Knowledge Graph-Driven Benchmark for Knowledge Conflict Detection

Anonymous ACL submission

Abstract

Knowledge conflict often occurs in retrieval-augmented generation (RAG) systems, where retrieved documents may be inconsistent with each other or contradict the model's parametric knowledge. Existing benchmarks for knowledge conflict detection have notable limitations, including a narrow focus on the question answering (QA) setup, heavy reliance on entity substitution techniques, and a limited range of conflict types. To address these gaps, we propose a knowledge graph (KG)-based data construction framework for knowledge conflict detection, ensuring greater diversity, complexity, and interpretability by leveraging the explicit relational structure of KGs. Experimental results on the new benchmark provide intriguing insights into the inner workings of LLMs in relation to knowledge conflict. They show that both open-source and proprietary LLMs struggle with conflict detection, particularly in multi-hop reasoning, and often fail to pinpoint the exact source of contradictions. These findings highlight the need for more robust benchmarks and improved methodologies for enhancing LLM reliability in conflict-aware reasoning.

1 Introduction

Retrieval-augmented generation (RAG) has become the de facto standard technique for enhancing the performance of large language models (LLMs), particularly in terms of updating their outdated knowledge and adapting to specialized domains (Lewis et al., 2020). While effective, its heavy reliance on the quality of retrieval always poses inherent risks. For instance, knowledge obtained from external sources may conflict with a model's parametric knowledge or even exhibit inconsistencies within the retrieved documents themselves.

Knowledge conflict is a recent research topic that covers issues related to the aforementioned cases and has been gaining consistent attention in the field (Xu et al., 2024). An ideal LLM-based system is expected to be robust in managing these challenges, generating reliable responses regardless of contradictory facts in its supporting data. However, its implementation is largely hindered by the difficulty of detecting whether disagreements exist across different knowledge sources and, if so, precisely where they occur.

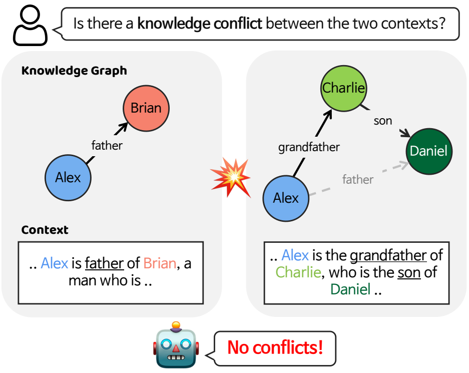

Is there a **knowledge conflict** between the two contexts?
Knowledge Graph
Brian
father
Alex
Context
.. Alex is father of Brian, a man who is ..
Charlie
son
grandfather
Daniel
father
Alex
.. Alex is the grandfather of Charlie, who is the son of Daniel ..
No conflicts!

Figure 1: Example of a knowledge graph-based multi-hop conflict in our benchmark, showing LLMs struggle to detect such cases requiring multi-hop reasoning.

Numerous benchmarks have been introduced to evaluate the performance of LLMs in knowledge conflict detection (Hsu et al., 2021; Li et al., 2024; Jiayang et al., 2024; Hou et al., 2024). Nonetheless, we emphasize that existing research on this subject has notable limitations. Firstly, previous studies have primarily focused on the question answering (QA) task, with conflicts occurring only between multiple answer candidates for a given question (Chen et al., 2022; Xie et al., 2024; Marjanovic et al., 2024). Secondly, prior research used overly simplistic techniques—e.g., entity substitution—for dataset construction (Longpre et al., 2021; Chen et al., 2022), which are insufficient to capture the complex nature of knowledge conflicts. Thirdly, while a few studies attempt to categorize conflict types—such as explicit versus implicit (Hou et al., 2024) and static, temporal, and disputable (Marjanovic et al., 2024)—there remains a lack of systematic analysis based on the classification of conflict types. Finally, with the rapid progress in the release of new, powerful LLMs, current benchmarks have become too straightforward to challenge sophisticated models—e.g., as we demonstrate in Section 5, GPT-4o-mini (OpenAI, 2024a) achieves over 80% accuracy on such datasets (Jiayang et al., 2024; Hou et al., 2024).

To alleviate these issues, in this work, we propose a framework for automatically constructing a benchmark for knowledge conflict detection, grounded in **knowledge graphs** (KGs). Specifically, our approach extracts subgraphs from a knowledge graph, each serving as the foundation of a knowledge chunk. Next, each subgraph is perturbed to generate variations, where modified nodes and edges in the perturbed graph introduce knowledge conflicts. Finally, both the original and perturbed graphs are transformed into text passages using KG-to-text algorithms powered by LLMs.

By its nature, the proposed method has several advantages. Since KGs provide a robust foundation for representing the relationships and structure of knowledge, our approach enables greater diversity, complexity, and controllability in inducing conflicts within documents (see Figure 1 for example). Furthermore, compared to text-based strategies, our method enhances the interpretability and structured analysis of the constructed dataset by visualizing conflicting entities and relations as graphs.

Lastly, we perform an extensive analysis using **Hierarchical Knowledge Conflict (HKC)**, a novel dataset for knowledge conflict detection developed through our framework. This benchmark features complex and varied conflict patterns, including multi-hop conflicts, which are rarely observed in previous benchmarks. Experiments on HKC reveal several insights into how current LLMs perceive knowledge conflicts, including: (1) both open-source and proprietary models remain imperfect at detecting conflicts, particularly when multi-hop reasoning is required; (2) even when models recognize contradictions, they struggle to pinpoint the exact position where the conflict occurs.

2 Related Work

Knowledge Conflict Knowledge conflict (KC) is a common issue in retrieval-augmented generation (RAG) systems, where retrieved documents may contain conflicting information or contradict the model's parametric knowledge.[1] To explore this phenomenon in greater depth, various datasets and configurations have been introduced.

Most studies focus on QA settings, presenting contradictory answers or evidence for given questions (Longpre et al., 2021; Chen et al., 2022; Xie et al., 2024; Jiayang et al., 2024). Entity-based methods (Chen et al., 2022) cause knowledge conflicts using various entity substitution techniques. Xie et al. (2024) enhances this by using LLMs to generate supporting evidence for conflicts.

Document-level conflict evaluates LLMs' ability to detect contradictions either across external documents (Jiayang et al., 2024; Hou et al., 2024; Marjanovic et al., 2024) or within a single document (Li et al., 2024), with the former being similar to our setting. Jiayang et al. (2024) constructs realistic evidence using LLMs, categorizing conflicts into answer and factoid conflicts. Hou et al. (2024) classifies conflicts as explicit or implicit based on Wikipedia's contradiction tags, while Marjanovic et al. (2024) distinguishes between static and dynamic facts using Wikipedia's edit logs.

Despite these efforts, we argue that previous benchmarks are not sufficiently challenging—e.g., they do not require multi-hop reasoning. Moreover, evaluations of knowledge conflict detection mostly focus on whether LLMs can identify conflicts rather than pinpointing their exact location. This work addresses these gaps by introducing a new dataset and evaluation metrics.

Data Construction with Knowledge Graphs Knowledge graphs play a crucial role in effectively solving various tasks—such as fact verification (Kim et al., 2023), question answering (Chen et al., 2024), and RAG (Sanmartin, 2024)—by providing structured knowledge representations.

KGs can also be actively leveraged for dataset creation. For instance, Meng et al. (2022) introduce COUNTERFACT, a dataset designed to assess factual modifications in transformer models. In contrast, this study utilizes KGs for the automatic construction of knowledge conflict benchmarks, a novel approach to the best of our knowledge.

3 Hierarchical Knowledge Conflict

Knowledge graphs provide a powerful framework for structurally representing knowledge through en-

[1] KC is divided into 3 types based on the source of knowledge (Xu et al., 2024). We focus on inter-context conflicts arising from contradictions between two external documents.

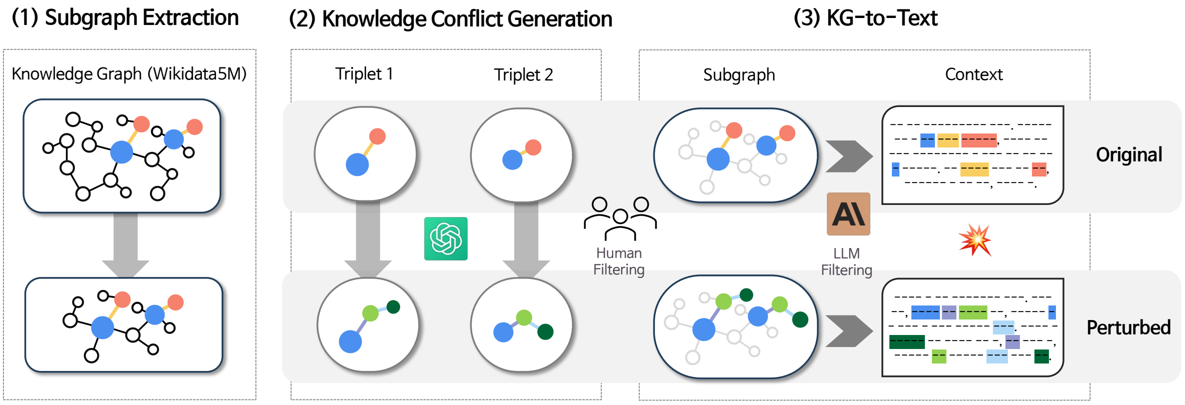


Figure 2: An overview of the proposed dataset construction framework.

tities and their relationships. This characteristic makes KGs particularly well-suited for constructing knowledge conflict datasets. By modifying nodes and edges within KGs, diverse and complex conflict patterns can be systematically introduced.

Another advantage of using knowledge graphs is that, unlike most existing benchmarks derived from QA datasets, they support broader domain coverage beyond specific tasks. Furthermore, as conflicts are not constrained by certain questions, KG-based conflicts exhibit a wider range of contradiction patterns than previous benchmarks.

In this study, we introduce a new benchmark, **Hierarchical Knowledge Conflict (HKC)**, constructed using our KG-based approach. The process comprises three steps, as depicted in Figure 2. First, subgraphs are extracted from a KG based on predefined criteria, acting as conceptual knowledge chunks (Section 3.1). Next, perturbations are applied to the subgraphs to provoke knowledge conflicts (Section 3.2). Finally, both the original and modified graphs are converted into text passages using KG-to-text algorithms (Section 3.3).

3.1 Subgraph Extraction

As the first step, we distill parts of a large-scale knowledge graph to build knowledge segments that serve as targets for inducing knowledge conflicts. Theoretically, any knowledge graph can be used as a source of information; in this work, we employ Wikidata5M (Wang et al., 2021). Wikidata5M consists of approximately 20 million triplets, covering various domains and knowledge structures.[2]

The key stages of subgraph extraction include seed triplet selection, graph traversal, and enforcing structural constraints.

Seed Triplet Selection We randomly sample seed triplets that function as the root for subgraph construction. Since they define the topic and structure of subgraphs, we filter relations involved in the seed triplets to facilitate the generation of hierarchy-based contradictions. Of the 825 unique relations in Wikidata5M, we use 24 selected relations.[3]

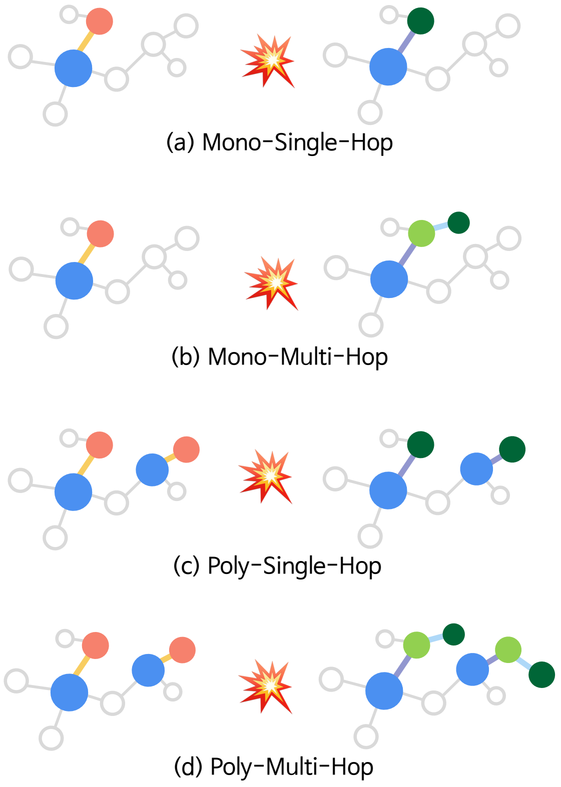
Graph Traversal Given the seed triplets, we perform graph traversal in the base KG, starting from the subject entity of each seed. We employ the Depth-First Search (DFS) algorithm to traverse the KG, progressively expanding the subgraph. DFS explores deep structural variations within the KG.

Enforcing Structural Constraints To maintain a balanced level of complexity and contextual richness in the extracted subgraphs, we impose the following constraints during DFS traversal:

- The number of edges in any extracted subgraph is capped at 15 to preserve computational feasibility and interpretability.
- To prevent excessive connections, we limit the number of edges per node to 5. This allows subgraphs to retain diverse structures without being dominated by a few highly connected nodes.
- The algorithm's maximum traversal depth is randomly decided for each traversal, resulting in subgraphs of diverse diameters and structures.

[2] For the diversity and robustness of the final dataset, we preprocess the KG as follows. Entities hard to be functionally defined, e.g., emoticons and special symbols (4,000 in total), are removed. In addition, general concepts and nodes with too many connections—e.g., 'human' and 'United States'—are excluded. The 30 most connected nodes are filtered out.

[3] See Appendix A.1 for the full list of the used relations.



(a) Mono-Single-Hop

(b) Mono-Multi-Hop

(c) Poly-Single-Hop

(d) Poly-Multi-Hop

Figure 3: Four different types of conflicts in HKC.

3.2 Knowledge Conflict Generation

In this phase, the goal is to perturb and modify the originally extracted subgraphs to create counterparts that contain information contradicting the original. To this end, we leverage advanced LLMs with deep reasoning capabilities, anticipating that they can generate reasonable and creative candidates that introduce knowledge conflicts within a given context. Nonetheless, the naïve use of such models does not guarantee success in this task, as they are inherently imperfect at recognizing knowledge conflicts. We thus propose a method to guide LLMs in reliably generating contradictory facts.

Category of Conflicts As illustrated in Figure 3, we consider four different types of knowledge conflicts in dataset generation. These categories are defined by the number of hops in reasoning paths involved in conflicts and the total number of conflicts present in a graph. **Single-Hop** conflicts (Figure 3 (a) and (c)) arise when a conflict occurs within a single triplet, affecting either a node or an edge. **Multi-Hop** conflicts (Figure 3 (b) and (d)) display conflicts spanning multiple triplets. Finally, based on the number of conflicts in a subgraph, they can be classified as **Mono**-type (a single conflict) or **Poly**-type (two or more conflicts).

Knowledge Conflict Generation Prompt

Instruction

You will be provided with a triplet set, [Original Triplet]. Modify it to introduce a **multi-hop knowledge conflict** with [Original Triplet].

# REQUIREMENTS
- The conflict must contradict the original triplet.
- The conflict must span at least **two or more hops** instead of directly contradicting the original triplet.
- The response should include the modified triplet **along with the intermediate steps** leading to the conflict.
- Do not include any explanation.

Here are a example of Multi-Hop Conflict Generation:

Example

[ORIGINAL TRIPLET] (tocantins (state) | divides into | novo jardim)
[MODIFIED TRIPLET] (tocantins (state) | borders | mato grosso) (mato grosso | contains | novo jardim)

Figure 4: Prompt for generating multi-hop conflicts.

Single-Hop		Multi-hop		Total
Mono	Poly	Mono	Poly	
92	41	59	42	234

Table 1: Data statistics for HKC by conflict category.

Triplet-Level Prompting We use OpenAI's o3-mini (OpenAI, 2025) to collect conflict candidates. Specifying a subgraph as a set of subject-relation-object triplets, we sample the target triplet for perturbation.[4] We then prompt the LLM to generate new triplet candidates that contradict the original.

Still, we found that naïve prompting is insufficient for generating diverse and logically complex conflicts.[5] As a solution, we provide an example of real conflicts to encourage the model to consider more than simple entity or relation swaps. The final prompt template used in this process is depicted in Figure 4. Note that it is designed for multi-hop conflicts, and a prompt for single-hop conflicts can be readily derived from it. For Poly-type conflicts, we repeat the entire process with the same graph until the desired number of conflicts is obtained.

[4] Our initial attempts revealed that providing the entire subgraph context confuses the model, making the task impossible. Therefore, we focus on triplet-level operation in this work.

[5] The model relied on repetitive patterns, struggling to generate creative conflicts and failing to produce knowledge conflicts that truly contradict common sense.

> **KG-to-Text Transformation Prompt**
>
> **Instruction**
>
> Your task is to write a brief, fluent, and coherent single-paragraph in natural language. The text should be balanced and neutral. Make sure that all the triplets mentioned in the text can be derived from the input data, do NOT add any extra information.

Figure 5: Prompt for KG-to-Text transformation.

Manual Review To ensure data quality, we conduct a human review. In-house researchers manually reviewed the data instances, removing any of poor quality. As a result, HKC comprises a total of 234 examples, with detailed statistics presented introduced in Table 1. Compared to existing datasets such as ECON and WikiContradict, mentioned in Section 4, our dataset contains a reasonable number of instances, especially considering the challenge of constructing difficult yet natural conflicts.

3.3 KG-to-Text Transformation

To represent knowledge conflicts from graphs in natural language, we finally apply KG-to-text transformation. We largely follow the approach outlined in Kasner and Dusek (2024), using the prompts specified in Figure 5 to guide GPT-4o-mini (OpenAI, 2024a) in generating coherent textual contexts while preserving the meaning of the input subgraph. For data quality control, we also perform automatic verification using Claude-3.5-Sonnet (Anthropic, 2024b), with the prompt shown in Figure 12.

4 Experimental Setups

We conduct conflict detection experiments using the dataset we constructed. The goal of the tested models is to detect contradictions, if any, arising from discrepancies between two given documents.

We evaluate various open-source and proprietary LLMs for conflict detection without applying any task-specific training. Instead, we prompt them to recognize potential contradictions. The following paragraphs detail the LLMs, datasets, prompting strategies, and metrics used in our experiments.

LLMs We use 5 LLMs: Mixtral-8x7B Instruct (team, 2023), Llama 3.1 70B Instruct (Dubey et al., 2024), Claude 3 Haiku (Anthropic, 2024a), GPT-4o-mini (OpenAI, 2024a), o1 (OpenAI, 2024b).

> **Knowledge Conflict Detection Prompt**
>
> **Instruction**
>
> You are given two contexts and your goal is to determine if there are any conflicts between them. Ignore what you know and make a judgment.
> (1) If no conflicts exist, simply respond with 'No conflicts'.
> (2) If conflicts exist, count the existing conflicts and specify each conflict reason briefly, and (3) present those all conflicting sentence pairs from each context.

Figure 6: Prompt for Knowledge Conflict Detection.

Datasets Experiments include both existing benchmarks and our newly introduced dataset, creating a comprehensive evaluation framework.

- **ECON** (Jiayang et al., 2024): A dataset created by introducing evidence conflicts through two methods—answer conflicts and factoid conflicts—highlighting contradictions in supporting evidence. It contains 168 data instances.
- **WikiContradict** (Hou et al., 2024): A human-annotated QA benchmark utilizing Wikipedia's contradiction tags to capture real-world knowledge conflicts. It categorizes contradictions into explicit and implicit types. After deduplication, it comprises 103 data samples.
- **HKC**: The dataset proposed in this study is constructed based on knowledge graphs (KGs), with conflicts arising from their structure. It includes both single- and multi-hop conflicts, making it more diverse and complex than previous ones.

Prompting Strategy Prior work (Jiayang et al., 2024; Hou et al., 2024) usually formulates the problem as a binary classification (yes/no), using simple prompts for LLMs. In contrast, we adopt a stepwise prompting strategy (see Figure 6) to explore the maximum capability of LLMs for this task.

(1) **Identification:** LLMs are first prompted to recognize whether a knowledge conflict exists between the given passages.

(2) **Explanation:** If a conflict is detected, LLMs should explicitly justify why they believe conflicts exist, encouraging logical reasoning rather than relying solely on surface-level knowledge.

Models / Datasets	ECON	WikiContradict	HKC
Mixtral 8x7B	46.43	52.43	**15.45**
Llama 3.1 70B	81.41	78.79	**67.98**
Claude 3.5 Haiku	83.33	61.17	**57.08**
GPT-4o-mini	88.10	82.52	**78.11**
o1	74.40	74.76	**64.38**
Average	74.73	69.93	**56.60**

Table 2: Results (%) on three KC detection datasets, measured by **Conflict Identification (CI)**. Lower scores indicate greater difficulty for models.

Models / Datasets	ECON	WikiContradict	HKC
Mixtral 8x7B	60.26	77.78	**41.67**
Llama 3.1 70B	53.54	65.38	**29.75**
Claude 3.5 Haiku	74.29	85.71	**55.64**
GPT-4o-mini	68.92	83.53	**49.45**
o1	87.20	87.01	**76.67**
Average	68.84	79.88	**50.64**

Table 3: Results (%) on three KC detection datasets, measured by **Conflict Localization (CL)**. Lower scores indicate greater difficulty for models.

(3) **Localization:** LLMs are instructed to identify the exact sentences or statements where conflicts occur, assessing their ability to pinpoint the precise source of contradictions.

Metrics To consider the stochasticity of LLMs, all models perform three separate inference runs. We rely on two metrics for fine-grained evaluation. All metrics are averaged over all data instances in a dataset. Note that these scores are manually computed by participating researchers, as automatic evaluation methods, such as LLM-as-a-judge, are not yet reliable enough for this task.[6]

- **Conflict Identification (CI):** If a model fails to detect a conflict in any of the three attempts, it receives a score of 0; otherwise, it receives 1.
- **Conflict Localization (CL):** For cases where LLMs successfully detect a conflict (CI score = 1), we further evaluate their performance in conflict localization. LLMs must correctly identify all conflicting sentences within a given context to receive a score of 1; otherwise, they receive 0.

5 Experimental Results

The main experimental results are presented in Table 2 and Table 3. A lower score on a dataset indicates that LLMs struggle more with it, demonstrating its higher level of difficulty.

Overall Results LLMs tested on HKC consistently show lower CI and CL scores compared to those on ECON and WikiContradict, with average scores decreasing by up to 18% and 30%, respectively. This indicates that models struggle to identify conflicts in our dataset, and even when they do, they have difficulty pinpointing the exact portions where the conflict occurs.

Conflict Identification (CI) Scores per LLM From Table 2, we observe that, model-wise, GPT-4o-mini achieves the highest performance, while Mixtral consistently records the lowest. Mixtral's lowest score of 15% underscores its significant weakness in identifying conflicts, a trend also observed in previous studies. Llama exhibits a distinct trend, particularly with the HKC dataset, failing to provide an answer in 23.5% of the three inference attempts and frequently refusing to respond directly to queries. In contrast, GPT-4o-mini demonstrates strong conflict identification capabilities, achieving over 74% success on previous KC datasets and maintaining similar performance on ours, confirming its effectiveness across all datasets.

Conflict Localization (CL) Scores per LLM Table 3 shows a similar trend to Table 2, with o1 achieving the best performance and Llama the worst. Although o1 does not achieve the highest CI scores, once it detects a conflict, it demonstrates strong stepwise reasoning, effectively specifying the exact location of the conflict. It also produces the shortest and most concise responses. Conversely, Llama generates significantly longer responses than other models, suggesting it misclassifies non-conflicting sentences as conflicts, resulting in a substantial drop in CL.

Performance by Conflict Types Figure 7 shows the average performance of all LLMs, categorized by the four conflict types defined in Section 3.2. Single-hop conflicts are related to entity or relation substitutions, where models perform relatively well in both identification and localization. However, multi-hop conflicts introduce greater complexity, making contradictions more indirect and resulting in lower CI and CL scores. Particularly in localization, multi-hop conflicts become more challenging as they span across various locations.

Meanwhile, a higher number of conflicts indi-

[6] If it were possible, further investigation of knowledge conflict detection would be unnecessary.

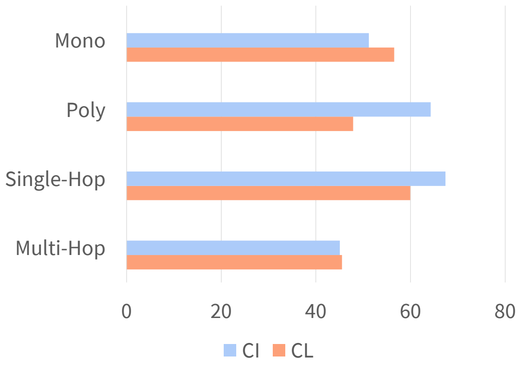


Figure 7: Average performance analysis of LLMs on HKC by conflict type. More conflicts aid recognition but hinder precise localization, with multi-hop cases being inherently more challenging than single-hop ones.

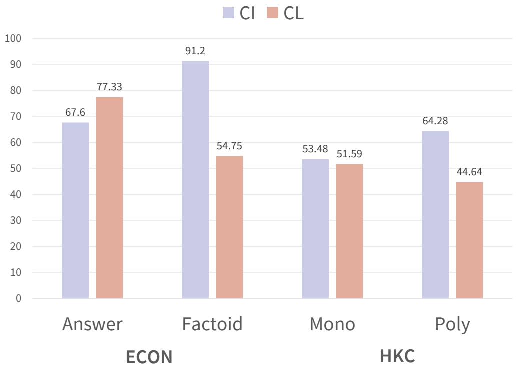


Figure 8: Comparison of detection performance according to the number of conflicts. ECON's factoid conflicts include multiple conflicts spanning across sentences.

cates a greater degree of contradiction between the two contexts, making it easier for models to detect conflicts. A similar trend was observed in ECON, where the CI score increased with the number of conflicts, as shown in Figure 8. Note that ECON's factoid conflicts involve multiple conflicts introduced across several sentences. This aligns with our findings, suggesting that while a higher number of conflicts facilitates conflict identification, it also makes precise localization more challenging. Conversely, when multiple conflicts occur, identifying all specific conflicting sentences becomes more difficult, leading to a decrease in the CL score.

6 Analysis

Classification of Data Instances from Existing KC Detection Datasets To analyze knowledge conflict patterns in previous datasets through our categorization (e.g., Poly-Multi-Hop), we apply our proposed conflict typology to existing KC datasets, namely ECON and WikiContradict. By representing these datasets as knowledge graphs, we assess how their knowledge conflicts align with our classification scheme.

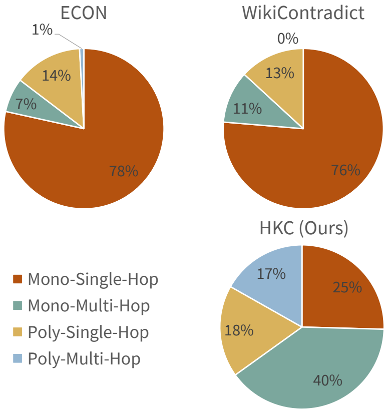


Figure 9: Proportions of four conflict types across three KC detection datasets. HKC demonstrates greater diversity and complexity than the other two datasets.

A challenge in analyzing existing KC datasets is the absence of a predefined ontology and domain structure. As a result, traditional ontology-based knowledge representation methods (van Cauter and Yakovets, 2024) are difficult to apply. To address this, we utilize the LangChain (Chase, 2022) framework to construct reliable knowledge graphs in schema-free environments, ensuring a structured and interpretable representation of knowledge conflicts.

Figure 9 presents the classification results of ECON and Wikicontradict datasets based on our typology. The results show that Mono-Single-Hop conflicts are the most prevalent type, with 78% in ECON and 76% in Wikicontradict. These types of conflicts are relatively easier to detect as they typically involve straightforward contradictions within a single document.

In contrast, our dataset (HKC) exhibits a much higher presence of Mono-Multi-Hop (39%) and Poly-Multi-Hop (18%) conflicts compared to existing datasets. These results reinforce the claim that our dataset presents more challenging knowledge conflict cases, demanding more sophisticated reasoning capabilities from LLMs.

Comparison of Difficult Conflict Types We compare a challenging subset of an existing KC dataset with those from our dataset. pecifically, we analyze *implicit* conflicts in WikiContradict and multi-hop conflicts in our dataset. WikiContradict

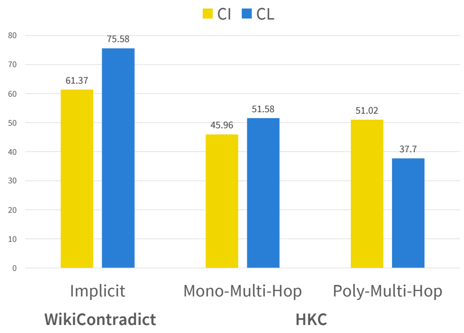


Figure 10: Comparison of the difficulty in challenging subsets of two KC datasets.

includes both explicit and implicit conflicts, with the latter being more challenging for models.

Figure 10 presents the CI and CL performance results for different conflict types. The results show that multi-hop conflicts in our dataset are more challenging to resolve than WikiContradict's implicit conflicts. CI performance for our dataset's multi-hop conflicts is up to 15% lower, while CL performance drops by up to 37% compared to WikiContradict's implicit conflicts. This highlights the increased complexity of our dataset, demanding more advanced conflict resolution abilities.

KG vs. Text-based Dataset Creation We compare the performance of knowledge conflicts generated using knowledge graphs (KG-based) and those created based on textual context (text-based). The goal is to assess which method produces more diverse and challenging conflicts.

We select Mono-Single-Hop and Poly-Single-Hop conflict data from HKC to evaluate each method's effectiveness in generating different types of conflicts. For a fair comparison, the text-based approach uses prompts adapted from the KG-based method, modified for a purely textual context.

Table 4 presents the CI scores for conflicts generated by each method. The results show that KG-based knowledge conflict generation produces more challenging conflicts than the text-based approach. Specifically, the KG-based method yields CI scores that are 11% lower for single-hop conflicts and 8% lower for multi-hop conflicts. This suggests that KG-based conflict generation is more effective in introducing difficult contradictions.

Effectiveness of Prompts for Conflict Detection Previous studies (Jiayang et al., 2024; Hou et al., 2024) on KC detection rely on binary (yes/no) prompts to determine conflict presence, making the detection process overly simplistic. In contrast, our research employs a multi-step prompt approach to enhance the accuracy of knowledge conflict detection (refer to Section 4).

	Text-based	KG-based
Mono-Single-Hop	92.73	**81.36**
Mono-Multi-Hop	78.82	**70.65**

Table 4: Comparison of CI scores between KG-based vs. context-based construction methods. Lower is better.

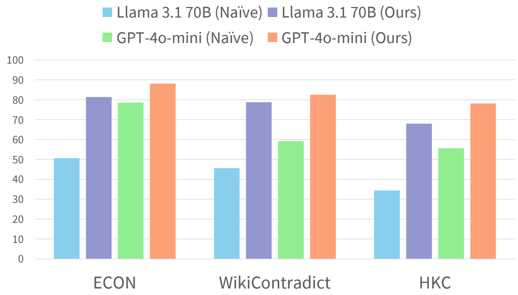


Figure 11: Comparison of the effectiveness of prompts for knowledge conflict detection, tested on two models.

To compare the effectiveness of different knowledge conflict detection prompts, we conducted conflict detection experiments on three datasets: ECON, Wikicontradict, and HKC. Figure 11 presents the results of these experiments.

The results show that across all datasets and models, our multi-step prompt outperforms the naive prompt. Specifically, our prompt achieves a minimum CI performance improvement of 9.5% and a maximum improvement of 33.56% over the naive prompt. These findings indicate that a multi-step approach to conflict detection is more effective, providing greater accuracy in identifying knowledge conflicts. These improvements highlight the importance of structuring the conflict detection process through detailed, multi-step reasoning rather than relying on simplistic binary prompts.

7 Conclusion

We propose a KG-based benchmark for knowledge conflict detection with greater diversity and complexity. Results on this dataset reveal the strengths and limitations of LLMs in handling knowledge conflicts. Despite recent progress, LLMs continue to struggle with conflict detection in complex cases, e.g., those equiring multi-hop reasoning. As a future direction, we aim to develop an optimized method to help models overcome these limitations.

Limitations

While our proposed dataset and benchmark offer significant improvements in knowledge conflict detection, several limitations remain. The current dataset consists of 234 instances, which is relatively limited, though our framework is designed for scalability, and future work will focus on expanding the dataset with a more diverse range of conflicts. Additionally, our dataset primarily includes conflicts involving two conflicting knowledge statements, whereas expanding to multi-source conflicts could better reflect real-world knowledge inconsistencies. Manual verification is currently required for conflict generation and evaluation, but automating this process through advanced LLM-based filtering or weak supervision methods could enhance efficiency and scalability. Furthermore, our dataset is generated using Wikidata-based knowledge graphs, and incorporating other structured sources such as DBpedia, YAGO, and domain-specific knowledge graphs could enhance robustness and applicability. Addressing these limitations in future work will help enhance the robustness, scalability, and applicability of knowledge conflict detection in large-scale AI systems.

References

Anthropic. 2024a. Claude 3.5 Haiku.

Anthropic. 2024b. Claude 3.5 Sonnet.

H. Chase. 2022. LangChain.

Hung-Ting Chen, Michael Zhang, and Eunsol Choi. 2022. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2292–2307, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ruirui Chen, Weifeng Jiang, Chengwei Qin, Ishaan Singh Rawal, Cheston Tan, Dongkyu Choi, Bo Xiong, and Bo Ai. 2024. Llm-based multi-hop question answering with knowledge graph integration in evolving environments. *arXiv preprint arXiv:2408.15903*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yufang Hou, Alessandra Pascale, Javier Carnerero-Cano, Tigran Tchrakian, Radu Marinescu, Elizabeth Daly, Inkit Padhi, and Prasanna Sattigeri. 2024. Wikicontradict: A benchmark for evaluating llms on real-world knowledge conflicts from wikipedia. *arXiv preprint arXiv:2406.13805*.

Cheng Hsu, Cheng-Te Li, Diego Saez-Trumper, and Yi-Zhan Hsu. 2021. WikiContradiction: Detecting Self-Contradiction Articles on Wikipedia . In *2021 IEEE International Conference on Big Data (Big Data)*, pages 427–436, Los Alamitos, CA, USA. IEEE Computer Society.

Cheng Jiayang, Chunkit Chan, Qianqian Zhuang, Lin Qiu, Tianhang Zhang, Tengxiao Liu, Yangqiu Song, Yue Zhang, Pengfei Liu, and Zheng Zhang. 2024. ECON: On the detection and resolution of evidence conflicts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7816–7844, Miami, Florida, USA. Association for Computational Linguistics.

Zdeněk Kasner and Ondrej Dusek. 2024. Beyond traditional benchmarks: Analyzing behaviors of open LLMs on data-to-text generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12045–12072, Bangkok, Thailand. Association for Computational Linguistics.

Jiho Kim, Yeonsu Kwon, Yohan Jo, and Edward Choi. 2023. KG-GPT: A general framework for reasoning on knowledge graphs using large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9410–9421, Singapore. Association for Computational Linguistics.

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

Jierui Li, Vipul Raheja, and Dhruv Kumar. 2024. ContraDoc: Understanding self-contradictions in documents with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6509–6523, Mexico City, Mexico. Association for Computational Linguistics.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063.

Sara Vera Marjanovic, Haeun Yu, Pepa Atanasova, Maria Maistro, Christina Lioma, and Isabelle Augenstein. 2024. DYNAMICQA: Tracing internal knowledge conflicts in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14346–14360, Miami, Florida, USA. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

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

OpenAI. 2024b. Hello o1. *OpenAI*.

OpenAI. 2025. Hello o3-mini. *OpenAI*.

Diego Sanmartin. 2024. Kg-rag: Bridging the gap between knowledge and creativity. *arXiv preprint arXiv:2405.12035*.

Mistral AI team. 2023. Mixtral of experts.

Zeno van Cauter and Nikolay Yakovets. 2024. Ontology-guided knowledge graph construction from maintenance short texts. In *Proceedings of the 1st Workshop on Knowledge Graphs and Large Language Models (KaLLM 2024)*, pages 75–84, Bangkok, Thailand. Association for Computational Linguistics.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021. KEPLER: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, Miami, Florida, USA. Association for Computational Linguistics.

A Appendix

A.1 Selected Relation Lists for Subgraph Extraction

- P22 (father), P25 (mother), P1038 (father-in-law), P1066 (student of), P183 (endemic to)

- P828 (has cause), P463 (member of), P176 (made by), P361 (part of), P3179 (territory overlaps)

- P551 (lived in), P150 (contains), P807 (separated from), P2789 (connects with), P740 (originates from)

- P1889 (different from), P179 (part of the series), P460 (equivalent to), P1382 (overlaps with)

- P527 (consists of), P1923 (participating team), P54 (member of team), P1542 (has result), P355 (subsidiary)

A.2 Prompts for KG-to-Text Verification

KG-to-Text Verification Prompt

Instruction

You are an expert KG-to-text error detection system. Your task is to understand structured triplet data and determine whether the given context contains errors based on the following criteria:

- Incorrect: The triplet contradicts the context.
- Not checkable: The triplet cannot be checked in the context.
- Misleading: The triplet is present but creates a misleading interpretation in the context.

Your response must be a single categorical value:

- "No error": If none of the above errors are present.
- "Yes error": If any of the above errors are present.

...

Figure 12: Prompt for KG-to-Text verification.