MAGIC: A Multi-Hop and Graph-Based Benchmark for Inter-Context Conflicts in Retrieval-Augmented Generation

Anonymous ACL submission

Abstract

Knowledge conflict often arises in retrieval-augmented generation (RAG) systems, where retrieved documents may be inconsistent with one another or contradict the model's parametric knowledge. Existing benchmarks for investigating the phenomenon have notable limitations, including a narrow focus on the question answering setup, heavy reliance on entity substitution techniques, and a restricted range of conflict types. To address these issues, we propose a knowledge graph (KG)-based framework that generates varied and subtle conflicts between two similar yet distinct contexts, while ensuring interpretability through the explicit relational structure of KGs. Experimental results on our benchmark, MAGIC, provide intriguing insights into the inner workings of LLMs regarding knowledge conflict: both open-source and proprietary models struggle with conflict detection—especially when multi-hop reasoning is required—and often fail to pinpoint the exact source of contradictions. We also provide a wide range of Finally, we present in-depth analyses that serve as a foundation for improving LLMs in integrating diverse, sometimes even conflicting, information.

1 Introduction

Retrieval-augmented generation (RAG) has become the de facto standard for enhancing the performance of large language models (LLMs) by enabling updates to outdated knowledge and facilitating adaptation to specialized domains (Lewis et al., 2020). While effective, RAG's heavy reliance on retrieval quality introduces inherent risks. For instance, knowledge obtained from external sources may conflict with the model's parametric knowledge or contain inconsistencies within the retrieved documents themselves.

Knowledge conflict (KC) is a recent research topic that covers issues related to the aforementioned scenarios and has been receiving attention in the field (Xu et al., 2024). An ideal LLM-based system should be robust against knowledge conflict, capable of integrating information from multiple sources—including those that may be contradictory—and ultimately generating reliable responses. However, its implementation is largely hindered by the challenge of detecting whether disagreements exist across different knowledge sources and, if so, pinpointing exactly where they occur.

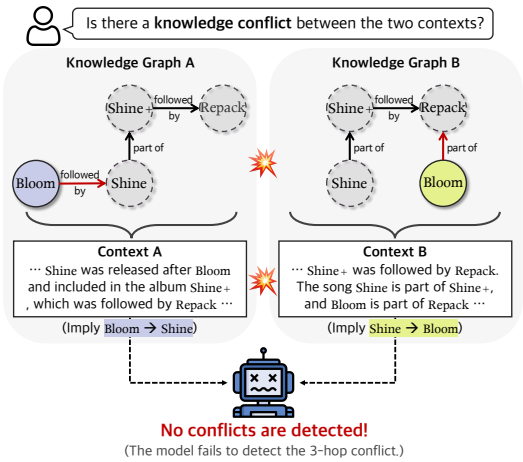

Figure 1: Example of a three-hop conflict from our benchmark, MAGIC. Even advanced LLMs struggle to detect subtle inconsistencies across two contexts, such as conflicting release orders of two songs.

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 in this area has notable limitations. First, previous studies primarily focus on the question answering (QA) task, where conflicts occur only among multiple answer candidates for a given question (Chen et al., 2022; Xie et al., 2024; Marjanovic et al., 2024). Second, prior research often relies on overly simplistic techniques for dataset construction, e.g., entity substitution (Longpre et al., 2021; Chen et al., 2022),

which are insufficient to capture the complex and subtle nature of knowledge conflicts. Third, while some studies attempt to categorize types of knowledge conflict (Hou et al., 2024; Marjanovic et al., 2024), systematic analysis distinguishing between forms, such as single-hop vs. multi-hop conflicts, is still lacking. Finally, existing benchmarks are largely concerned with exploring conflicts between parametric and external knowledge, while conflicts among multiple input documents remain underexplored (Jiayang et al., 2024; Hou et al., 2024).[1]

To alleviate these issues, we propose a framework for constructing a benchmark targeting ***inter-context knowledge conflict***. It leverages **knowledge graphs (KGs)** as the primary source, from which subgraphs are extracted to serve as the basis for distinct knowledge chunks. These subgraphs are then perturbed by modifying nodes and edges to introduce conflicts. Finally, both original and altered graphs are converted into corresponding text passages using KG-to-text generation algorithms.

By design, the proposed framework offers several advantages. Leveraging the relational structure of KGs enables greater diversity, complexity, and control in inducing conflicts within documents. Moreover, compared to text-based strategies, our approach improves interpretability and supports structured analysis by representing conflicting entities and relations in graph form (see Figure 1).

Lastly, we conduct extensive analyses using **MAGIC** (A **M**ulti-Hop **A**nd **G**raph-Based Benchmark for **I**nter-**C**ontext Conflicts), a novel dataset constructed through our framework. MAGIC features complex inter-context conflict patterns, including simultaneous and multi-hop cases rarely seen in existing benchmarks. Experiments on MAGIC yield key insights into how LLMs perceive knowledge conflicts: (1) most models remain imperfect at detecting conflicts, especially when multi-hop reasoning is required; and (2) even when contradictions are detected, models often fail to localize the exact point of conflict.

2 Related Work

Knowledge conflict benchmarks Early studies on synthetically inducing sentences with knowledge conflicts (KC) primarily rely on entity substitution (Chen et al., 2022), which is often subsequently polished and paraphrased using LLMs (Xie et al., 2024; Gokul et al., 2025). Regarding domains and tasks, most prior work focuses on potential inconsistencies arising in QA, such as contradictory answers provided for a given question (Longpre et al., 2021). However, this paradigm has a clear limitation: most conflicts must appear near answer-related contexts, which strictly limits the range of possible variations. To overcome this, we adopt question-free, KG-based methods, which inherently support more diverse problem types.[2]

Inter-context knowledge conflict detection evaluates an LLM's ability to identify contradictions either across multiple input contexts (Jiayang et al., 2024; Hou et al., 2024; Marjanovic et al., 2024) or within a single document (Li et al., 2024). Existing benchmarks follow two main strategies: collecting conflicts curated from Wikipedia (Hou et al., 2024) and generating artificial ones using LLMs (Jiayang et al., 2024). The former offers realism but limited coverage, while the latter is scalable yet often less natural and less representative of real-world scenarios. We aim to combine the strengths of both by leveraging the factual grounding of KGs and the generative fluency of LLMs, ensuring both high quality and scalability. Finally, we highlight that little work has examined performance with respect to fine-grained conflict types. Departing from the common convention of using only two conflict categories,[3] we organize our dataset by the number of conflicts and the reasoning hops required for resolution, enabling more systematic analysis.

KG-based dataset creation KGs play a crucial role in diverse tasks, e.g., fact verification (Kim et al., 2023), QA (Chen et al., 2024), and RAG (Sanmartin, 2024), by providing structured representations of knowledge. In addition, KGs can serve as valuable resources for dataset construction. For instance, Meng et al. (2022) introduce COUNTERFACT, a dataset designed to evaluate factual consistency and modifications in LLMs. In this paper, we also utilize KGs as a foundation for generating realistic and nuanced conflict statements. Bi et al. (2024) shares similarities with our work, as it also employs Wikidata triplets to induce

[1] Knowledge conflict is commonly studied in three setups (Xu et al., 2024): (1) *context–memory conflict*, involving inconsistencies between parametric and external knowledge; (2) *inter-context conflict*, where contradictions exist among multiple input documents; and (3) *intra-memory conflict*, where inconsistent responses reflect variation in an LLM's training data. In this work, we focus on ***inter-context conflict***.

[2] Concurrently, Gokul et al. (2025) propose a query-free method but rely solely on LLMs, unlike our KG-based one.

[3] Jiayang et al. (2024): answer & factoid; Hou et al. (2024): explicit & implicit; Marjanovic et al. (2024): static & dynamic.

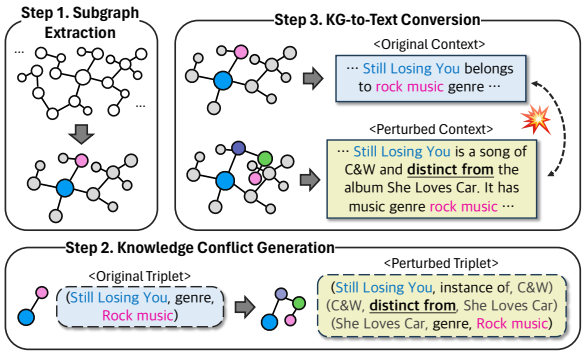

Figure 2: Overview of the proposed KG-based framework for benchmarking inter-context knowledge conflict detection. It comprises three steps: (1) Subgraph Extraction, (2) Knowledge Conflict Generation, and (3) KG-to-Text Conversion, with details listed in Section 3.

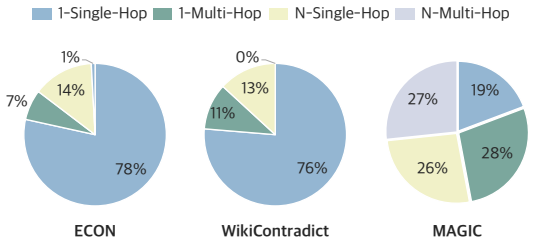

Figure 3: Distribution of conflict types across three knowledge conflict detection datasets. MAGIC demonstrates greater diversity and complexity than the others.

knowledge conflicts. However, their approach is limited to retrieving seed entities for substitution rather than leveraging the full structure of KGs.

3 MAGIC: Multi-Hop and Graph-Based Benchmark for Inter-Context Conflicts

In this section, we introduce a new framework for constructing an inter-context knowledge conflict detection dataset through the collaboration of KGs and LLMs. Compared to existing benchmarks, the proposed approach offers several advantages: (1) Rather than using QA datasets as the source, we adopt KGs, enabling broader domain coverage and a richer range of conflict forms, including multi-hop cases. (2) It combines the strengths of both manual and automated strategies by incorporating a human-in-the-loop process during LLM-based generation, ensuring both high quality and scalability. (3) It also enhances interpretability for users by providing two complementary views, one from the graph and the other from the corresponding text.

The procedure consists of three steps, as depicted in Figure 2. First, subgraphs are extracted from a KG based on predefined criteria, acting as conceptual knowledge chunks (§3.1). Next, perturbations are applied to subgraphs to provoke conflicts (§3.2). Finally, both original and modified graphs are converted into text passages using KG-to-text algorithms (§3.3). As a result, we present a benchmark named **MAGIC** (A **M**ulti-Hop **A**nd **G**raph-Based Benchmark for **I**nter-**C**ontext Conflicts).

3.1 Subgraph Extraction

As the first step, we distill parts of a large-scale KG to build knowledge segments that serve as targets for inducing knowledge conflicts. Theoretically, any KG can be utilized; in this work, we employ Wikidata5M (Wang et al., 2021). Wikidata5M consists of approximately 20 million triplets, covering various domains and knowledge structures.[4]

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 form the basis for subgraph construction. Since these triplets define the topic and structure of the resulting subgraphs, we filter the relations they contain. Specifically, from the 825 unique relations in Wikidata5M, we select 46 based on two criteria: (1) semantic clarity, which allows for controlled conflict manipulation, and (2) the ability to support meaningful multi-hop reasoning chains. To facilitate more detailed analysis, we group the selected relations into seven semantic domains based on their meaning and typical usage.[5]

Graph traversal Given the seed triplets, we perform graph traversal in the base KG starting from the subject entity of each seed. We use the Depth-First Search (DFS) algorithm to progressively expand the subgraph. DFS is well-suited for exploring deep structural variations within the KG.

Enforcing structural constraints We regulate DFS traversal with the following constraints to preserve both structural complexity and contextual coherence in the extracted subgraphs.

[4]For the diversity and robustness of MAGIC, we preprocess Wikidata5M 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.

[5]The domains are: Human, Geography, Organization, Creative Work, Class/Concept, Cause-Effect, and General. See Appendix A for the full list of relations used in each domain.

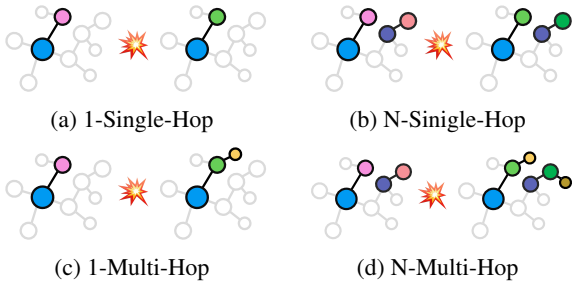

Figure 4: Four distinct types of conflicts in MAGIC.

- The number of edges in each extracted subgraph is capped at 15 to ensure computational feasibility and maintain interpretability.
- To prevent excessive connectivity, we limit the number of edges per node to 5. This ensures subgraphs retain structural diversity without being dominated by a few highly connected nodes.
- The maximum traversal depth of DFS is randomly determined for each run, resulting in subgraphs with diverse diameters and structures.

3.2 Knowledge Conflict Generation

In the second phase, the goal is to perturb and modify extracted subgraphs to create counterparts that contradicts the original. To this end, we leverage LLMs with strong reasoning capabilities, expecting them to generate plausible and creative candidates that introduce knowledge conflicts within a given context. However, naïvely using such models does not guarantee success, as they are inherently imperfect at recognizing knowledge conflicts. We therefore propose a method to guide LLMs in reliably generating contradictory facts.

Category of conflicts As illustrated in Figure 3, prior benchmarks, i.e., ECON (Jiayang et al., 2024) and WikiContradict (Hou et al., 2024), primarily target simple *1-Single-Hop* conflicts (Figure 4(a)) arising from individual facts. However, real-world discrepancies often involve multi-hop reasoning and multiple conflicts. To bridge this gap, we define eight distinct conflict types along two axes: (1) the number of reasoning hops required (**Single-Hop** vs. **Multi-Hop**) and (2) the number of conflicts present across the two contexts (**1** vs. **N**).

- **Single-Hop** conflicts (Figure 4(a), (b)) arise from inconsistencies through a single relation.
- **Multi-Hop** conflicts (Figure 4(c), (d)) require reasoning over multiple connected triplets.

Conflict Type	Single-Hop				Multi-Hop				Total
# Conflict	1	2	3	4	1	2	3	4	
# Instances	208	154	80	50	300	158	80	50	1,080

Table 1: MAGIC dataset statistics by conflict category.

Each conflict type is further divided into **1-conflict** and **N-conflict** cases, based on the number of contradictions observed between two given contexts, yielding four distinct types illustrated in Figure 4. The numbers of data instances allocated to each category is shown in Table 1. This categorization enables fine-grained evaluation across varying reasoning depths and conflict complexities.

Subgraph-level few-shot prompting We use OpenAI's o3-mini (OpenAI, 2025) to induce and collect conflict candidates. Given a target seed triplet, we prompt the LLM with both the triplet and its surrounding context, represented as a set of subject–relation–object triplets from the subgraph. Including the surrounding subgraph allows the model to generate natural contradictions that remain consistent with the local context.[6]

Still, we observe that prompting without task demonstrations (i.e., real examples of knowledge conflicts) often falls short in generating diverse and logically complex conflicts, particularly in multi-hop scenarios.[7] As a solution, we adopt a few-shot prompting strategy, where the prompt includes three validated conflict examples per relation type. This encourages the model to move beyond simple entity or relation swaps. The final prompt template used in this process is shown in Figure 12, Appendix B.1.[8] For N-conflict cases, the same graph is reused with different perturbations until the desired number of conflicts is achieved.

Quality control via human feedback Since LLMs are not inherently good at detecting knowledge conflicts, it may seem paradoxical to ask them to generate such cases. To alleviate this, we incorporate a human-in-the-loop process at two stages—before and after few-shot prompting—to improve dataset quality. First, human experts intervene during few-shot demonstration selection, manually filtering model-generated cases. Second, experts fil-

[6] Appendix F reveals that using only the target triplet often produced contextually misaligned conflicts.

[7] The model tends to repeat patterns, struggling to produce varied conflicts and often misaligning with common sense.

[8] While designed for multi-hop conflicts, the template can be readily adapted to single-hop settings.

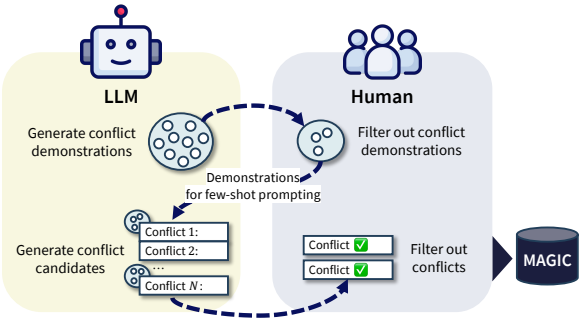

Figure 5: Two-stage human-in-the-loop pipeline for data quality control.

ter out trivial or incoherent outputs after conflict generation.[9] Figure 5 depicts the entire workflow.

3.3 KG-to-Text Conversion

To represent knowledge conflicts from graphs in natural language, we apply KG-to-text conversion, following the approach of Kasner and Dusek (2024) with modifications. Using the prompts shown in Figure 13, GPT-4o-mini (OpenAI, 2024a)[10] generates coherent textual contexts while preserving the semantics of the original graph. To ensure transformation accuracy, we perform automatic verification using Claude 3.7 Sonnet (Anthropic, 2025), with the prompt shown in Figure 14.

In addition, we validate data integrity by sampling and manually inspecting a subset of generated instances. Our manual inspection confirms that the conversion process achieves consistently high quality; further details are provided in Appendix B.2.

4 Experimental Setups

With the MAGIC benchmark as a foundation, we conduct experiments to examine how LLMs handle inter-context knowledge conflicts. We evaluate various open-source and proprietary LLMs without task-specific training, instead prompting them to identify potential contradictions. In the following, we outline the LLMs, datasets, prompting strategies, and metrics used in our experiments.

LLMs We use 5 LLMs: Llama 3.1 70B Instruct (Dubey et al., 2024), o1 (OpenAI, 2024b), Mixtral-8x7B Instruct (MistralAI, 2023), Claude 3.5 Haiku (Anthropic, 2024), GPT-4o-mini (OpenAI, 2024a).

Datasets Alongside MAGIC, we employ two existing benchmarks to highlight its strengths.

- **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.
- **MAGIC**: It is constructed atop KGs, where conflicts are systematically induced from the underlying relational structure. It encompasses both single-hop and multi-hop contradictions, with the number of conflicts dynamically varying across context pairs. In total, MAGIC consists of 1,080 carefully curated examples, with comprehensive statistics presented in Table 1. Remarkably, the scale of MAGIC surpasses that of existing benchmarks by a significant margin, offering a richer and more challenging resource for evaluating inter-context conflict detection.

Prompting strategy Prior work (Jiayang et al., 2024; Hou et al., 2024) typically frames the task as a binary classification problem, relying on minimal prompts (see Appendix B.3). In contrast, we introduce a stepwise prompting strategy (Figure 16) designed to more fully probe the capabilities of LLMs for inter-context conflict detection.

(1) **Identification:** LLMs determine whether a conflict exists between the given passages.

(2) **Explanation:** If so, they specify how many conflicts are present and explain why, encouraging logical reasoning beyond surface cues.

(3) **Localization:** LLMs pinpoint the exact sentences where conflicts occur, assessing their ability to locate contradiction sources.

Metrics To account for the stochasticity of LLMs, all models perform three independent inference runs. We use two metrics for fine-grained evaluation, with scores averaged across all instances in each dataset. These metrics are manually computed by participating researchers, as automatic methods—such as LLM-as-a-judge—remain insufficiently reliable for this task.[11]

[9] Two researchers independently reviewed disjoint subsets using a shared guideline. See Appendix E for full details.

[10] Specific version: gpt-4o-mini-2024-07-18.

[11] If they were, further investigation into knowledge conflict detection would be unnecessary.

Models / Datasets	ECON	WikiContradict	MAGIC
Mixtral 8x7B	46.43	52.43	**37.92**
Llama 3.1 70B	81.41	78.79	**73.83**
Claude 3.5 Haiku	83.33	61.17	**31.94**
GPT-4o-mini	88.10	82.52	**83.61**
o1	74.40	74.76	**68.06**
Average	74.73	69.93	**59.07**

Table 2: **ID** scores (%) on three KC detection datasets. Lower scores indicate higher task complexity.

Models / Datasets	ECON	WikiContradict	MAGIC
Mixtral 8x7B	35.71	40.78	**17.40**
Llama 3.1 70B	54.49	51.52	**37.89**
Claude 3.5 Haiku	66.07	52.43	**22.04**
GPT-4o-mini	63.69	68.93	**55.00**
o1	64.88	65.48	**49.72**
Average	57.09	55.74	**36.42**

Table 3: **LOC** scores (%) on three KC detection datasets. Lower scores indicate higher complexity. We observe that models struggle more with pinpointing the exact location of a conflict than with detecting its presence.

- Identification (**ID**) score: If a model fails to detect a conflict in any of the three attempts, it receives a score of 0; otherwise, it receives 1.
- Localization (**LOC**) score: We further evaluate LLMs' performance on conflict localization. A full score (1) is awarded only if all conflicting locations are correctly identified; otherwise, the score is 0. Note that this fine-grained evaluation has not been considered in previous work.

5 Experimental Results

The main experimental results are shown in Table 2 and Table 3. Lower scores indicate greater difficulty, suggesting the dataset is more challenging.

Overall results LLMs tested on MAGIC consistently show lower ID and LOC scores compared to those on ECON and WikiContradict, with average scores decreasing by up to 15% and 19%, respectively. This indicates that models struggle more to identify conflicts in our dataset, and even when they do, they have difficulty pinpointing the exact portions where the conflict occurs.

ID scores per LLM Table 2 shows that GPT-4o-mini achieves the highest accuracy and generalizes well to MAGIC. Mixtral performs the worst overall, while Haiku drops notably on MAGIC, suggesting difficulty with multi-hop reasoning. Llama shows a unique trend on MAGIC: it fails to respond in 47.4% of runs but achieves high ID scores when it does. Finally, o1 exhibits moderate performance, implying that conflict detection may depend on more than just LLMs' reasoning ability.

LOC scores per LLM Table 3 shows trends similar to Table 2, with GPT-4o-mini consistently achieving the best LOC scores across all datasets. Qualitative analysis reveals that o1 often takes a conservative stance, frequently predicting *no conflict* in ambiguous cases,[12] which leads to missed conflicts—particularly those requiring multi-hop thinking (see Table 8).[13] Mixtral's low LOC score largely stems from poor initial conflict identification, reducing the number of evaluated localization cases. Llama shows moderate performance but produces overly long outputs (1121.5 tokens vs. 631.1 for others), often including irrelevant content. This suggests over-reliance on instructions, leading to over-detection and reduced localization precision.

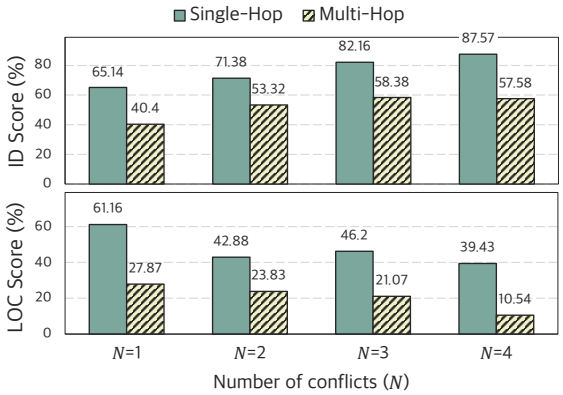

Figure 6: Average LLM performance on MAGIC by conflict type. A greater number of conflicts aids recognition but hinders localization, while multi-hop cases remain inherently more challenging than single-hop.

Performance by conflict types Figure 6 presents the average performance of all LLMs across four conflict complexity settings: single-hop vs. multi-hop and 1-conflict vs. N-conflict, as in Section 3.2. Detailed results are provided in Appendix D.1.

Single-hop conflicts, often involving entity or relation substitutions, are relatively easier, with models performing well in both identification and local-

[12] e.g., (A consists of B) vs. (A has C), (B not part of C).

[13] This outcome is unexpected, given o1's strong reasoning abilities in math and coding. One possible explanation is the exclusion of explicit reasoning instructions such as "think step by step". We leave further investigation on this as future work.

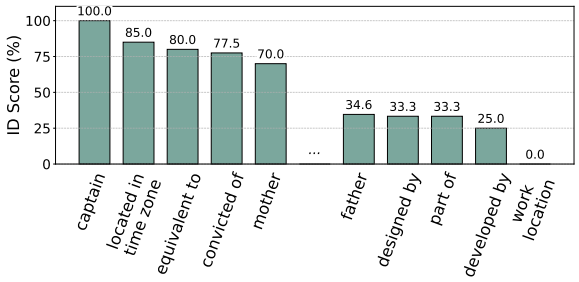

Figure 7: Average performance of LLMs on the 5 most predictive and 5 least predictive relations. This analysis focuses soly on the single-conflict subset of MAGIC.

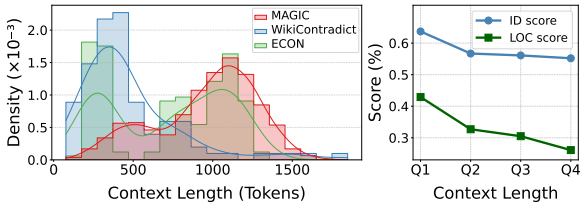

Figure 8: (Left) Context length distributions of the three KC datasets. (Right) ID and LOC scores averaged over LLMs decline as context length increases in MAGIC.

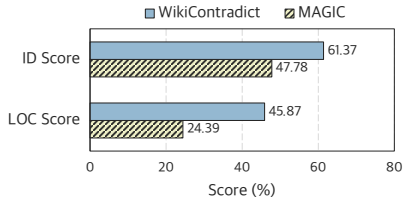

Figure 9: Comparison of difficulty between challenging subsets of WikiContradict and MAGIC, with scores averaged over all five LLMs.

ization. In contrast, multi-hop conflicts introduce greater complexity, as contradictions become more indirect, leading to lower ID and LOC scores. Localization is especially difficult in multi-hop cases, as conflicts often span multiple locations.

Meanwhile, a higher number of conflicts reflects a stronger contradiction between the two contexts, making conflict detection easier for models.[14] In other words, while more conflicts facilitate identification, they also complicate precise localization. When multiple conflicts occur, pinpointing all specific conflicting sentences becomes more difficult, leading to lower LOC scores.

Domain- & relation-level analysis As MAGIC spans diverse topics and relations in KGs, it enables analysis at both the domain and relation levels.[15] In the domain-specific evaluation, we find that most LLMs perform relatively well on the Class/Concept domain, likely because it contains clear and well-defined relations (e.g., *subclass of*, *different from*) that are readily recognizable.[16] In contrast, performance on the Organization domain varies widely—from around 75% (GPT-4o-mini, Llama) to below 15% (Haiku, Mixtral)—indicating that models differs in handling hierarchical relations.

Meanwhile Figure 7 highlights the top five most and least predictive relations across all LLMs. Conflicts involving *captain* and *mother* are relatively easy for models, while *work location* and *father* pose greater challenges. More representative examples are provided in Table 13 in the Appendix.

Finally, Figure 19 shows that detection performance is influenced by the number of domains present in the context. In multi-domain cases, ID scores tend to improve—likely due to increased semantic diversity making conflicts more salient. In contrast, LOC scores decline, possibly because structural variation complicates span localization.

6 Discussion

Taxonomy-based analysis To systematically analyze conflict patterns, we apply our proposed conflict typology to ECON and WikiContradict. Representing these datasets as graphs allows us to highlight their characteristics. A key challenge is the lack of predefined ontologies or domain structures, which hinders the use of traditional ontology-based methods (van Cauter and Yakovets, 2024). To address this, we use LangChain (Chase, 2022) to construct reliable, schema-free KGs that support structured conflict representation.

Figure 3 shows that 1-Single-Hop conflicts—the typically easiest case—are the most prevalent in prior datasets, accounting for 78% in ECON and 76% in WikiContradict. In contrast, MAGIC exhibits a balanced distribution across conflict types, with substantial proportions of 1-Multi-Hop (28%) and N-Multi-Hop (27%), underscoring its robustness as a benchmark.

Length-based analysis In conflict detection, a reasonable hypothesis is that context length positively correlates with dataset difficulty, as longer documents tend to involve more complex linguistic structures. To validate this, we present length-

[14] A similar trend is also observed in ECON (Jiayang et al., 2024). See Appendix G for details.

[15] Detailed domain-level statistics are reported in Table 9.

[16] Still, exceptions exist depending on the used model.

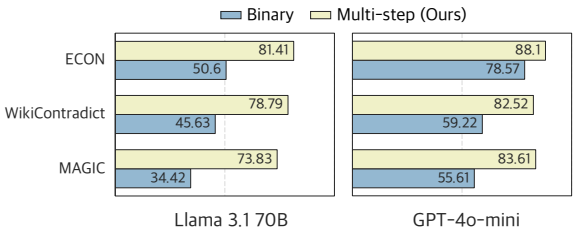

Figure 10: Comparison of ID scores for binary and multi-step prompts, tested on two models.

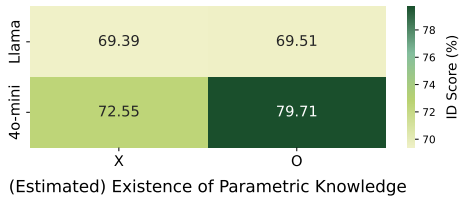

Figure 11: ID scores of two LLMs on known (O) vs. unknown (X) instances, based on their parametric knowledge. Models demonstrate stronger conflict detection when the relevant knowledge is already embedded.

oriented anlysis in Figure 8. The left panel visualizes the total context length distribution among the three datasets, with MAGIC containing more long-context examples. We suspect MAGIC's multi-hop and description-rich design contributes to this.

Further, we group MAGIC into four bins based on context length and report their respective performance in the right panel of Figure 8. To balance group sizes, we use quantile-based binning: Q1 contains the shortest contexts, and Q4 the longest. As context length increases, both scores decline, with LOC dropping more sharply—indicating the growing difficulty of pinpointing conflicting spans in longer inputs.

Comparison of challenging subsets We conduct a comparative study between the challenging subset of WikiContradict (Implicit) and the corresponding subset from our dataset (Multi-Hop in MAGIC).[17] Figure 9 reports that MAGIC proves more difficult for LLMs than WikiContradict. MAGIC yields an ID score up to 15% lower and a LOC score up to 23% lower compared to WikiContradict. This highlights MAGIC's intrinsic complexity, setting it apart from existing datasets.

Impact of prompts While prior work (Jiayang et al., 2024; Hou et al., 2024) typically employs binary (yes/no) prompts for conflict detection—often oversimplifying the task—we use a multi-step technique that helps improve performance. To compare the effectiveness of these two distinct prompting strategies for conflict detection, we conduct experiments, with results reported in Figure 10. Across all cases, the results show that our multi-step prompting approach outperforms the naïve prompt, achieving improvement up to 39.41%. This implies that although we employ a method superior to those commonly used in the literature, there remains substantial room for addressing the challenges posed by MAGIC.

[17]WikiContradict consists of explicit and implicit conflicts; the latter are considered more challenging due to their subtlety.

Impact of parametric knowledge In the literature on *inter-context* conflict, research has primarily focused on conflicts between two input documents, overlooking the influence of parametric knowledge which can significantly affect the performance of LLMs in knowledge conflict detection.

To address this, we explore an underexamined setup by splitting a subset of MAGIC—instances with 1- and 2-conflicts—into two groups based on the estimated presence of parametric knowledge in LLMs. Concretely, we approximate the existence of parametric knowledge for a given triplet by posing a converted verification question—for example, 'Is Barack Obama Sr. the father of Barack Obama? We label a triplet as known if the model provides the correct answer in at least 4 out of 5 attempts, and as unknown if it succeeds in no more than 1.

As shown in Figure 11, both GPT-4o-mini and Llama 3.1 achieve higher ID scores on known instances. These findings suggest that models may be more effective at detecting conflicts when they already possess the relevant factual knowledge. While this approach does not provide a rigorous measure of internal knowledge, it offers a coarse-grained perspective. We remain a more systematic investigation for future work.

7 Conclusion

We propose a KG-based benchmark, MAGIC, for inter-context knowledge conflict detection with greater diversity and complexity. Experimental results 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 requiring multi-hop reasoning. As a future direction, we aim to develop an optimized method to help models overcome these limitations.

Limitations

While MAGIC offers a novel benchmark for evaluating knowledge conflict detection, particularly inter-context conflict, it still presents several areas for improvement. First, MAGIC is constructed using Wikidata-based knowledge graphs. Incorporating additional sources—such as DBpedia, YAGO, or domain-specific knowledge graphs—could enhance its robustness and broaden its applicability. A promising direction is to align semantically equivalent relations across these graphs to ensure consistency and coverage. In addition, localization evaluation currently relies on human judgment in this work, developing more automated approaches could enable more fine-grained evaluation in future work. 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. 2024. Claude 3.5 Haiku.

Anthropic. 2025. Claude 3.7 Sonnet.

Baolong Bi, Shaohan Huang, Yiwei Wang, Tianchi Yang, Zihan Zhang, Haizhen Huang, Lingrui Mei, Junfeng Fang, Zehao Li, Furu Wei, et al. 2024. Context-dpo: Aligning language models for context-faithfulness. *arXiv preprint arXiv:2412.15280*.

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*.

Vignesh Gokul, Srikanth Tenneti, and Alwarappan Nakkiran. 2025. Contradiction detection in rag systems: Evaluating llms as context validators for improved information consistency. *arXiv preprint arXiv:2504.00180*.

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.

Mistral AI. 2023. Mistral 7B.

MistralAI. 2023. Mixtral of experts.

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*.

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 Details of Selected Relation Lists per Domain

We selected a subset of relations from Wikidata5M based on two criteria: (1) semantic clarity, which enables controlled manipulation, and (2) the ability to form meaningful multi-hop reasoning chains, essential for constructing complex conflict scenarios.

To facilitate more detailed analysis and structured subgraph extraction, we grouped the selected relations into seven semantic domains—*Human, Geography, Organization, Creative Work, Class/Concept, Cause-Effect*, and *General*. This categorization was manually determined by referring to the official description and subject type constraint of each property on its Wikidata page, based on their inherent meaning and typical usage patterns.

The complete list of selected relations and their domains is shown in Table 4.

Domain	Relations
Human	P22 (father), P25(mother), P551 (lived in), P634 (captain), P937 (work location), P1038 (father-in-law), P1066 (student of), P1344 (participant in), P1399 (convicted of), P737 (inflenced by)
Geography	P36 (capital), P47 (shares border with), P150 (contains), P189 (find location), P197 (next station), P421 (located in time zone), P1336 (territory claimed by), P3179 (territory overlaps), P1382 (overlaps with), P2789 (connects with)
Organization	P127 (owned by), P463 (member of), P807 (separated from), P1001 (belongs to), P2652 (partnership with)
Creative Work	P144 (based on), P155 (follows), P178 (developed by), P264 (record label), P287 (designed by)
Class/Concept	P279 (subclass of), P460 (equivalent to), P461 (opposite of), P1889 (different from)
Cause-Effect	P828 (has cause), P1478 (has immediate cause), P1479 (has contributing factor), P1537 (contributing factor of), P1542 (has result)
General	P361 (part of), P527 (consists of), P1011 (excluding), P2283 (uses), P3094 (develops from), P4330 (contains)

Table 4: Selected relations from Wikidata5M.

B Prompts used in MAGIC

B.1 Prompt for Knowledge Conflict Generation

Figure 12 shows a prompt used to generate multi-hop conflicts. It introduces constraints that guide the LLM to construct indirect contradictions using related subgraphs. The prompt also defines a clear output format and encourages the use of specific entities and relations in the surrounding subgraphs. Section 3.2 mentions how to select few-shot demonstrations used during generation. For N-conflict cases, the same graph is reused with different perturbations to create multiple conflicts.

Knowledge Conflict Generation Prompt

Instruction

You will be provided with an [Original Triplet] and a set of [Related Subgraphs]. Your task is to generate a **multi-hop knowledge conflict** consisting of 2–3 logically connected triplets that together create a logically coherent but indirect conflict with the [Original Triplet].

REQUIREMENTS:
- Construct a conflict that does not directly contradict the original triplet, but introduces contradiction through intermediate reasoning steps.
- Use **one or more specific entities or relations** from the [Related Subgraphs] to build the multi-hop chain.
- Each triplet must be semantically valid and form a realistic knowledge path.
- The conflict must be concrete, not vague or overly inferential.

OUTPUT FORMAT:
Return a set of 2–3 triplets in the form (Subject | Relation | Object) that together form the multi-hop conflict. Do not include explanations, reasoning steps, or any additional text.

Demonstrations

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

Figure 12: Prompt for generating multi-hop conflicts.

B.2 Prompt for KG-to-Text Conversion

We include two prompts used in our KG-to-text pipeline. Figure 13 prompt guides the generation of natural language contexts from input subgraphs, and Figure 14 performs automatic verification of triplet coverage using Claude 3.7 Sonnet. Only the outputs that return *No error* in this verification step are retained in our dataset to ensure high-quality generation.

To further guarantee the trustworthiness of the model-based verification, we conduct human inspection on 167 sampled outputs spanning all data

Criteria	Coverage (%)
Conflict Triplet	95.21
Subgraph Triplet	82.04

Table 5: Human inspection results of KG-to-text outputs based on conflict expression and subgraph coverage.

types in our dataset. Human annotators evaluate each context using a two-step protocol:

- **Conflict Triplet Coverage:** Does the text include the *target* or *perturbed* triplet (i.e., is the intended conflict expressed)?
- **Subgraph Triplet Coverage:** Does the text include *all* subgraph triplets (i.e., is the overall information faithfully conveyed)?

As shown in Table 5, our method achieves 95.21% accuracy in the first criterion and 82.04% in the second.[18] These results demonstrate that our KG-to-text pipeline is both reliable and automatable, ensuring high-quality generation with minimal manual intervention.

KG-to-Text Conversion Prompt

Instruction

Your task is to convert every provided triplet into a brief, fluent, natural, and coherent single-paragraph in natural language. You MUST include all the facts from the provided triplets. Do NOT omit any triplet or add any information that is not present in the triplets, even if it seems plausible or more natural.

Figure 13: Prompt for KG-to-Text Conversion.

B.3 Prompt for Knowledge Conflict Detection

Figure 15 shows the binary prompt used in (Jiayang et al., 2024) for knowledge conflict detection. In contrast, our stepwise conflict detection prompt is shown in Figure 16, as mentioned in Section 4.

C Linguistic quality of MAGIC

Since the contexts in MAGIC are generated by LLMs, assessing their linguistic quality is important to ensure their naturalness and realism. To this end, we conducted an evaluation using Claude 3.7 Sonnet. For a fair comparison, we sampled 155 contexts from each dataset (WikiContradict, ECON, MAGIC) to match overall length distributions. Each context was rated on a 0~5 scale, and the evaluation prompt is provided in Figure 17 and 18.

As shown in Table 6, MAGIC is only 0.32 lower in naturalness than WikiContradict, despite the latter being based on real Wikipedia text. It also surpasses ECON in realism, indicating our dataset generates plausible, real-world-like scenarios. These results demonstrate that MAGIC serves as a reliable benchmark for evaluating knowledge conflict detection.

[18] With subgraphs often involving more than 10 triplets, achieving over 80% coverage indicates consistent preservation of essential information.

KG-to-Text Verification Prompt

Instruction

You are an expert KG-to-text error detection system. Your task is to verify whether the provided context faithfully reflects the given set of triplets. You must identify any errors based on the following criteria:

- INCORRECT: The triplet contradicts factual information stated in the context.
- NOT CHECKABLE: The triplet is not verifiable because the necessary information is missing from the context.
- MISLEADING: The triplet appears to be present but introduces a misleading or confusing interpretation in the context.

Response Format

Your response must be one of the following two values only:

- "NO ERROR": if none of the above errors are present, and the paragraph is concise and fluent.
- "YES ERROR": if any of the above errors are present, or if the paragraph is unnaturally verbose or lacks fluency.

Provide only the answer without any additional explanation.

Figure 14: Prompt for KG-to-Text Verification.

Conflict Detection (Binary Prompt)

Do the two pieces of context contain conflicting information on answering the question? (Yes/No)

Figure 15: Binary prompt for conflict detection.

Conflict Detection (Multi-step Prompt)

Instruction

You are given two contexts and your goal is to determine if there are any factual conflicts between them. Ignore what you know and only consider the information within the two contexts.

Response Format

If there are no conflicts, output: `No conflicts`
If there are conflicts, output in this exact format:
Conflicts: `<number_of_conflict>`
Conflict 1:
- Reason: `<description_of_conflict>`
- Sentence A: `"<sentence_from_context A>"`
- Sentence B: `"<sentence_from_context B>"`
.. (Repeat for each conflict)

Figure 16: Multi-step prompt for conflict detection.

Context Naturalness Evaluation Prompt

Evaluate the naturalness of the following context on a scale from 0 (very poor) to 5 (excellent).

Focus on:

- Is the context grammatically correct?
- Does it sound fluent and stylistically natural, as if written by a human?

Do not consider factual accuracy or whether the content is realistic.
Return only a single integer from 0 to 5. Do not provide any explanation or reasoning.

Figure 17: Prompt for evaluating naturalness of context.

D Detailed Model Evaluation on MAGIC

D.1 Model Performance by each Conflict Type

Table 7 and Table 8 provide detailed ID and LOC scores for five models across different conflict types in MAGIC dataset. Table 7 includes smaller model performnce, such as Mistral-7B Instruct-v0.1 (Mistral AI, 2023) and Llama 3.1 8B Instruct (Dubey et al., 2024). As discussed in Section 5, when the number of conflicts (N) grows, ID scores tend to increase while LOC scores decrease. Multi-hop conflicts generally yield lower performance in both ID and LOC compared to single-hop cases.

D.2 Domain-Level Analysis

Table 9 presents the average ID scores of models across seven domains, and Figure 7 shows the averaged ID scores of models across relation types.

Context Realism Evaluation Prompt

Evaluate the realism of the following context on a scale from 0 (very poor) to 5 (excellent).

Focus on:

- Could this context plausibly occur in a real-world setting?
- Does it resemble something that could realistically appear in natural use cases?

Do not consider grammar or fluency. Also, do not check whether it is factually accurate.
Return only a single integer from 0 to 5. Do not provide any explanation or reasoning.

Figure 18: Prompt for evaluating realism of context.

Metric	ECON	WikiContradict	MAGIC
Naturalness	4.36	4.39	4.08
Realism	4.00	4.72	4.26

Table 6: Naturalness and realism scores of KC datasets.

These results are under the 1-conflict setting in MAGIC. For each model, the highest score is highlighted in red, while the lowest is in green.

Figure 19 shows the performance based on the number of domains included in each data instance.

D.3 Comparison with existing KC datasets

In addition to Section 6, we also compare our MAGIC dataset with existing datasets by evaluating 1-Single-Hop type across all datasets to ensure a fair comparison. Even in the relatively simple 1-Single-Hop conflicts, Figure 20 shows that models perform worse on our dataset than on ECON and WikiContradict, suggesting that our data introduces challenges beyond surface-level contradiction.

E Annotation Guideline

To ensure the quality of our MAGIC dataset, human intervention was applied at two stages of data construction pipeline: (1) manually selecting of few-shot demonstrations used for prompting, and (2) filtering out trivial or inherent model-generated conflicts after generation.

All annotations were performed independently by two researchers following a shared guideline to ensure consistency across relation types and conflict settings. The detailed annotation guidelines for each stage are provided below.

	Single-Hop				Multi-Hop			
	N=1	N=2	N=3	N=4	N=1	N=2	N=3	N=4
Mixtral 8x7B	42.72	51.66	51.90	67.35	23.47	31.61	38.16	30.00
Llama 3.1 70B	72.43	79.31	93.88	90.48	59.52	78.67	70.00	73.91
Claude 3.5 Haiku	38.94	48.05	67.50	84.00	12.33	13.92	26.25	28.00
GPT-4o-mini	85.58	87.66	100.00	98.00	70.67	84.18	86.25	94.00
o1	87.02	90.26	97.50	98.00	37.00	58.23	71.25	62.00
Avg.	65.14	71.38	82.16	87.57	40.40	53.32	58.38	57.58
Mistral 7B	7.43	11.82	7.41	10.81	11.06	6.93	4.00	12.90
Llama 3.1 8B	7.21	15.59	23.00	24.00	6.33	13.38	10.00	12.00

Table 7: ID Score by Model on MAGIC.

	Single-Hop				Multi-Hop			
	N=1	N=2	N=3	N=4	N=1	N=2	N=3	N=4
Mixtral 8x7B	38.83	21.85	15.19	14.29	12.59	7.10	6.58	0.00
Llama 3.1 70B	62.64	42.53	40.82	42.86	31.75	25.33	25.00	8.70
Claude 3.5 Haiku	38.94	29.22	47.50	34.00	9.67	8.86	12.50	8.00
GPT-4o-mini	78.85	53.90	58.75	44.00	54.67	47.47	33.75	24.00
o1	86.54	66.88	68.75	62.00	30.67	30.38	27.50	12.00
Avg.	61.16	42.88	46.20	39.43	27.87	23.83	21.07	10.54

Table 8: LOC Score by Model on MAGIC.

Domain	Mixtral	Llama	Haiku	GPT-4o-m	o1
Human	22.64	56.52	32.73	69.09	63.64
Geography	25.95	72.73	19.50	79.87	56.60
Organization	14.52	75.86	9.52	74.60	42.86
Creative Work	37.93	66.67	17.24	82.76	41.38
Class/Concept	47.89	68.97	33.78	75.68	72.97
Cause-Effect	32.56	54.55	25.58	72.09	53.49
General	42.86	47.22	25.88	78.82	56.47

Table 9: Domain-level analysis with 1-conflict problems in MAGIC.

Guideline for Few-shot Demonstration Selection

The goal of this task is to select three three representative examples per relation type to be included in the few-shot prompt. These examples should be chosen from zero-shot model generations and must serve as effective demonstrations of plausible and challenging knowledge conflicts.

Selected examples should follow the criteria below:

- Each example must express a plausible and semantically coherent knowledge conflict.
- The conflict must be appropriate for the given relation, reflecting its intended usage in Wikidata.
- Examples involving multi-hop reasoning or indirect contradictions were preferred over surface-level entity substitutions.
- Redundant or structurally repetitive patterns across examples were avoided to ensure diversity.

Guideline for Post-generation Conflict Filtering

The goal of this task is to identify and remove low-quality outputs from model-generated conflict instances. Annotators should review each instance generated via prompting and apply the following criteria to filter out unsuitable samples:

- For single-hop conflicts, the perturbed triplet must contradict the original triplet in a direct and unambiguous manner.
- For multi-hop cases, the contradiction must emerge through a reasoning chain spanning multiple triplets.
- In N-conflict instances, each conflict must be logically independent and non-overlapping with others in the same context pair.
- Outputs with unnatural phrasing, semantic incoherence, or implausible context were discarded.

F Impact of Subgraph Scope on Conflict Generation

In a preliminary attempt, we prompted the model to generate conflicts using only a selected seed triplet, without incorporating the surrounding subgraph. Table 10 shows that this often resulted in trivial patterns—such as simply negating the original relation or replacing with incoherent, off-topic facts. These observations underscore the importance of incorporating broader context, as our subgraph-level prompting enables the generation of more realistic

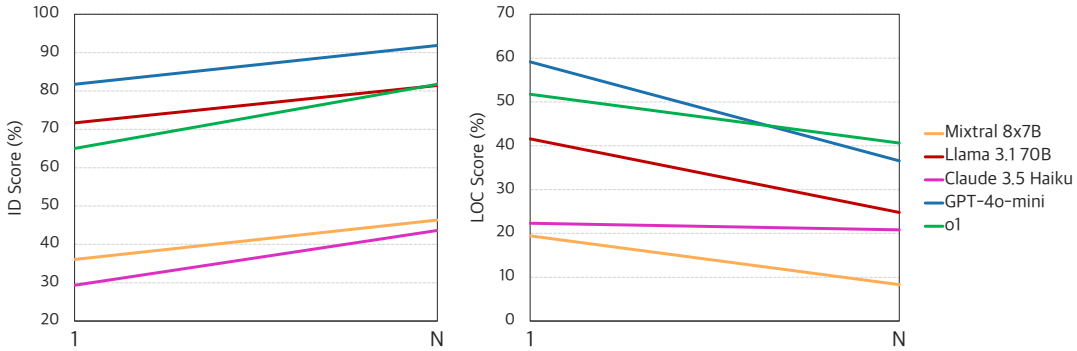

Figure 19: Performance by Number of Domains per Sample in MAGIC.

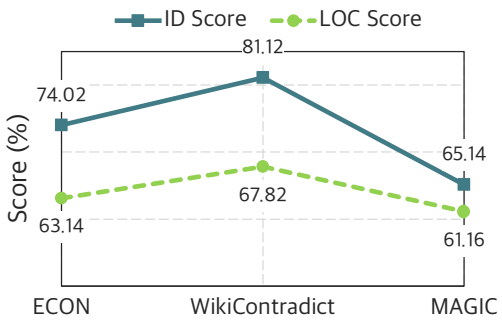

Figure 20: Comparison of detection performance in 1-Single-Hop conflict on three datasets.

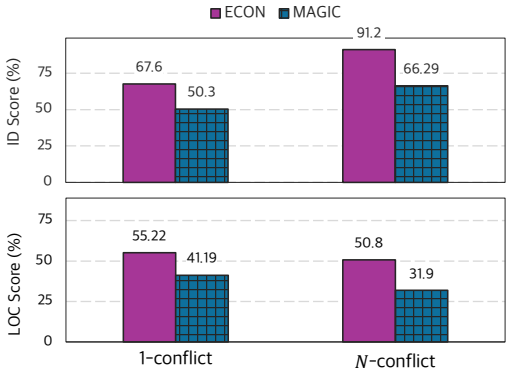

Figure 21: Comparison of detection performance by number of conflicts. ECON's factoid conflicts contain multiple (N) conflicts that span across sentences.

and semantically grounded conflicts.

G Performance by Number of Conflicts

Figure 21 compares performance across ECON and MAGIC based on the number of conflicts (1-conflict vs. N-conflict). 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 LOC score.

H Examples from MAGIC Dataset

Table 11 and 12 show example contexts from MAGIC dataset.

H.1 Qualitative Analysis of Model Failures

Beyond domain-level trends, individual failures reveal deeper challenges. GPT-4o-mini, despite its overall strength, struggles with multi-hop reasoning over densely connected subgraphs. As shown in the below of Table 13, one MAGIC includes a case where *John is equivalent to Hans*, while the perturbed ones ultimately imply that *John is not equivalent to Hans*. Detecting this contradiction requires chaining equivalence and distinction relations across multiple entities.

1-Multi-Hop	
Original Triplet	(Moskva | contains | staroye kryukovo district)
Perturbed Triplet	(Moskva | borders | Odintsovo), (Odintsovo | contains | staroye kryukovo district)
Subgraph	(Moskva | contains | kosino-ukhtomsky district), (Moskva | contains | Prospekt Vernadskogo District), (Moskva | divides into | Chertanovo Tsentralnoye District), (Moskva | twinned administrative body | tunis)
Context1	Moskva is a city that contains several districts, including the staroye kryukovo district, kosino-ukhtomsky district, and Prospekt Vernadskogo District. Additionally, it is divided into Chertanovo Tsentralnoye District. Moskva also has a twinned administrative body relationship with Tunis.
Context2	Moskva borders Odintsovo and contains several districts, including the Kosino-Ukhtomsky District, the Prospekt Vernadskogo District, and it also divides into Chertanovo Tsentralnoye District. Additionally, Odintsovo contains the Staroye Kryukovo District, and Moskva is twinned with the administrative body of Tunis.
2-Single-Hop	
Original Triplet #1	(Hastings, New Brunswick | territory overlaps | Kings County, New Brunswick)
Perturbed Triplet #1	(Hastings, New Brunswick | territory does not overlap | Kings County, New Brunswick)
Original Triplet #2	(Hastings, New Brunswick | territory overlaps | albert county)
Perturbed Triplet #2	(Hastings, New Brunswick | territory does not overlap | albert county)
Subgraph	(Hastings, New Brunswick | instance of | a dark-sky preserve) (Hastings, New Brunswick | operator | canadian parks service), (Hastings, New Brunswick | member of | Canadian National Parks), (Canadian National Parks | subclass of | national park), (Canadian National Parks | has list | List of national parks of Canada), (Canadian National Parks | subclass of | Protected areas of Canada), (albert county | located in the administrative territorial entity | Culture of New Brunswick), (albert county | shares border with | saint john county)
Context1	Hastings, New Brunswick, is a dark-sky preserve operated by the Canadian Parks Service and is a member of Canadian National Parks, which is a subclass of national parks and protected areas in Canada. The territory of Hastings overlaps with Kings County and Albert County, the latter of which is located within the administrative territorial entity of the Culture of New Brunswick and shares a border with Saint John County.
Context2	Hastings, New Brunswick, is recognized as a dark-sky preserve and is operated by the Canadian Parks Service, making it a member of the Canadian National Parks, which is a subclass of national parks and protected areas in Canada. The territory of Hastings does not overlap with Kings County or Albert County, the latter of which is situated within the administrative territorial entity of the Culture of New Brunswick and shares a border with Saint John County. The Canadian National Parks also maintains a list known as the List of national parks of Canada.

Table 10: Example of Triplet-level Dataset Generation.

1-Single-Hop	
Context1	Guy Williams, a basketball player, is distinct from Gus Williams, another basketball player. He has been a member of the Baltimore Bullets and the Oakland Warriors during his career. It is important to note that he is different from James “Fly” Williams. The name Guy, which is his given name, is equivalent to the name Guido and belongs to the French vocabulary. Additionally, the name Guy is also a surname that is identical to the given name. The writing system used for the name Guy is Latin alphabet letters, which are based on the roman-alphabet and are an instance of an alphabetic writing system. The history of the Latin alphabet is the historical context surrounding the use of these letters.
Context2	Guy Williams, a basketball player, is the same person as Gus Williams, who is also known as a basketball player. The name Guy is equivalent to the given name Guido, is of French vocabulary origin, and shares a family name identical to Guy (surname). The writing system for the name Guy is Latin alphabet letters, which are instances of an alphabetic writing system based on the Roman alphabet. Latin alphabet letters have a historical context in the history of the Latin alphabet. In his basketball career, Guy Williams was a member of the Baltimore Bullets and the Oakland Warriors, and he is different from another player named James “Fly” Williams.
2-Single-Hop	
Context1	The concept of the "Margin of opportunity" overlaps with both the "Sensitive period" and the "Time limit," and is classified as a subclass of the broader category of "event." This margin is also a facet of both "WikiProject Urban studies" and "Orbital maneuver." Within the realm of knowledge, "mastery" is seen as a subclass of "Knowledgeableness" and "aptitude," and is described by the source known as "el panson." Additionally, the term "event" is used by a "Relativistic observer" and is equivalent to "Event (statistics)." Notably, "WikiProject Urban studies" itself falls under the subclass of "mastery."
Context2	The Margin of Opportunity is disjoint from the Sensitive Period and does not overlap with the Time Limit. It is considered a subclass of events and a facet of both WikiProject Urban Studies and Orbital Maneuver. Mastery is a subclass of both Knowledgeableness and Aptitude, and is described by the source El Panson. The concept of an event is used by a relativistic observer and is equivalent to an event in statistics. Lastly, WikiProject Urban Studies is a subclass of mastery.
3-Single-Hop	
Context1	Bilecik University is a public college located in Bilejik and has separated from several institutions, including Kütahya Dumlupınar University, Eskişehir Osmangazi University, and Anadolu University.
Context2	Bilecik University is a public college located in Bilejik and has merged with several institutions, including Kütahya Dumlupınar University, Eskişehir Osmangazi University, and Anadolu University.
4-Single-Hop	
Context1	John, a personal name, is equivalent to Ifan in some contexts and can also be represented as Jean and Ioannis in different languages. In addition, the name John has a specific connection to John Hervey, who lived from 1616 to 1680 and was the father-in-law of Robert Jermyn. John Hervey was involved in significant historical events, as he was a member of the Royal Society of Great Britain and the Restoration Parliament, and he lived during the Civil War in England. The name John is primarily associated with the German language (iso 639:deu), which is classified as a High German dialect, has a V2 word order, possesses a simple will future tense, and includes grammatical cases such as the genitive case.
Context2	John is a personal name that is not equivalent to the given name Ifan, but it is equivalent to the names Hans and Johannes. The German language, denoted as iso 639:deu, is a High German language characterized by v2 word order, the simple will future ii tense, and the genitive case. John Hervey, who lived from 1616 to 1680, was the son-in-law of Robert Jermyn and was involved in the Civil War in England. He was also a member of the Royal Society of Great Britain and the Restoration Parliament.

Table 11: Examples for Single-Hop conflict in MAGIC.

1-Multi-Hop	
Context1	Hastings, New Brunswick, is an area that overlaps with Saint John County and Albert County, and it is recognized as a dark-sky preserve, an instance of a terrestrial protected area. This preserve, which is part of the Fundy Biosphere Reserve, is under the operation of the Canadian Parks Service and has been conferred the designation of a dark-sky preserve by both the Dark Sky Association and the Bulletin of the Royal Astronomical Society of Canada. The Dark Sky Association, an environmental organization focused on creating darker skies through initiatives like "lights out for darker skies," has its field of work centered on this ecological effort and is also located within the administrative territorial entity of Satori Charter School.
Context2	Hastings, located in New Brunswick, overlaps with Albert County and is part of the Fundy Biosphere Reserve, operated by the Canadian Parks Service. This area is recognized as a dark-sky preserve, a designation conferred by both the Bulletin of the Royal Astronomical Society of Canada and the Dark Sky Association, which is an environmental organization focused on promoting darker skies through initiatives like "Lights Out for Darker Skies." It is important to note that Albert County is completely disjoint from Saint John County.
2-Multi-Hop	
Context1	Auitzotl was the son of Atotoztli II and Epcoatl. He had a sibling named Acolnahuacatl Tzacualcoatl and was the father of two children, Chimalpilli II and Cuahatemoc.
Context2	Auitzotl is the parent of Cuahatemoc and Chimalpilli II, and has a sibling named Acolnahuacatl Tzacualcoatl, who is the child of Epcoatl. Cuahatemoc's mother is Atotoztli II.
3-Multi-Hop	
Context1	The State Penn is a member of several organizations, including the Digital Library Federation, SPARC Europe, and the Center for Research Libraries (CRL). Additionally, it is affiliated with the Oak Ridge Associated Universities, which is located in Oak Ridge, Tennessee, and operates as a matrix organization. Otto Poggeler, who serves as an employee at State Penn, was born in Attendorn, Germany, and is a member of the North Rhine-Westphalia Academy for Sciences and Arts. He speaks and writes in German, known by the ISO 639 code as deu.
Context2	State Penn is a member of Oak Ridge Associated Universities, which is a matrix organization located in Oak Ridge, Tennessee. The Digital Library Federation excludes membership for matrix organizations, and similarly, SPARC Europe restricts membership to institutions located in Europe. Oak Ridge Associated Universities is also considered a matrix organization, which is incongruent with the Center for Research Libraries (CRL). In addition, Otto Poggeler, who was born in Attendorn, Germany, speaks German and is a member of the North Rhine-Westphalia Academy for Sciences and Arts. He is employed by State Penn.
4-Multi-Hop	
Context1	The name "Iulian" is equivalent to several other names including "Julian," "Julio," "Juliusz," and "Julien." In Modern Spanish, the name "Julián" serves as its counterpart and is also equivalent to "Jules," "Julien," and "Juliusz," while "Iulian" further connects to "Julián" as a first name. Additionally, Modern Spanish is classified under the Castilian languages and features various grammatical moods and tenses, including the conditional tense, present indefinite tense, and past perfect simple.
Context2	The name Julián is a given name in Modern Spanish, equivalent to several other names including iulian, Juliusz, jules, and Julien, though it is distinct from the name julian. The origins of the name Julio can be traced back to Latin, while Julián's various equivalents reflect its connections across different languages and cultures. All references to Julián confirm its continuous usage in Modern Spanish, which is a subclass of Castilian languages characterized by grammatical moods such as the conditional tense and various tenses including the present indefinite and past perfect simple.

Table 12: Examples for Multi-Hop conflict in MAGIC.

Easy Example	
Original Triplet	(1891 British Lions tour to South Africa | captain | bill maclagen)
Perturbed Triplet	(1891 British Lions tour to South Africa | captain | william burrows)
Subgraph	(1891 British Lions tour to South Africa | destination point | Suid Africa), (1891 British Lions tour to South Africa | follows | 1888 british lions tour to new zealand and australia), (1891 British Lions tour to South Africa | sport | Fifteen-a-side), (1891 British Lions tour to South Africa | followed by | 1896 British Lions tour to South Africa), (Suid Africa | diplomatic relation | hellada), (Suid Africa | part of | Continent of Africa), (Suid Africa | diplomatic relation | argentina), (1896 British Lions tour to South Africa | followed by | 1899 british lions tour to australia), (1896 British Lions tour to South Africa | destination point | Suid Africa)
Context1	The 1891 British Lions tour to South Africa, captained by Bill Maclagen, was a fifteen-a-side rugby tour that took place in Suid Africa, following the 1888 British Lions tour to New Zealand and Australia. This tour was later followed by the 1896 British Lions tour to South Africa, which also had Suid Africa as its destination point. Suid Africa is located on the Continent of Africa and maintains diplomatic relations with Hellada and Argentina. The 1896 tour would subsequently be followed by the 1899 British Lions tour to Australia.
Context2	The 1891 British Lions tour to South Africa, captained by William Burrows, was a Fifteen-a-side rugby tour that took place in Suid Africa, following the earlier 1888 British Lions tour to New Zealand and Australia. This tour was succeeded by the 1896 British Lions tour to South Africa, which also had Suid Africa as its destination point. Suid Africa, part of the Continent of Africa, has diplomatic relations with Hellada and Argentina. The 1896 British Lions tour to South Africa was subsequently followed by the 1899 British Lions tour to Australia.
Difficult Example 1	
Original Triplet	(Henry Charles Fitzroy Somerset, 8th Duke of Beaufort | work location | united kingdom/london)
Perturbed Triplet	(Henry Charles Fitzroy Somerset, 8th Duke of Beaufort | place of death | Stoke Gifford), (Stoke Gifford | is part of | South Gloucestershire), (South Gloucestershire | is geographically distinct from | united kingdom/london)
Subgraph	(Henry Charles Fitzroy Somerset, 8th Duke of Beaufort | position held | uk mp), (Henry Charles Fitzroy Somerset, 8th Duke of Beaufort | place of death | Stoke Gifford), (Henry Charles Fitzroy Somerset, 8th Duke of Beaufort | member of political party | new tories), (Henry Charles Fitzroy Somerset, 8th Duke of Beaufort | child | Blanche Scott Douglas)
Context1	Henry Charles Fitzroy Somerset, the 8th Duke of Beaufort, worked in London, United Kingdom, and held the position of MP for the UK as part of the New Tories political party. He passed away in Stoke Gifford, and is also known to have had a child named Blanche Scott Douglas.
Context2	Henry Charles Fitzroy Somerset, the 8th Duke of Beaufort, passed away in Stoke Gifford, a locality that is part of South Gloucestershire, which is geographically distinct from London in the United Kingdom. He held the position of a UK Member of Parliament and was a member of the New Tories political party. He is also the father of Blanche Scott Douglas.
Difficult Example 2	
Original Triplet	(Jon (first name) | equivalent to | Hans (first name))
Perturbed Triplet	(Jon (first name) | equivalent to | gianni (first name)), (gianni (first name) | equivalent to | Ivan), (Ivan | different from | Hans (first name))
Subgraph	(gianni (first name) | equivalent to | jaan (first name)), (gianni (first name) | equivalent to | Ivan), (gianni (first name) | equivalent to | juan (first name)), (gianni (first name) | equivalent to | Ioannis (name)), (Jon (first name) | equivalent to | Evan), (Jon (first name) | equivalent to | jaan (first name)), (Jon (first name) | equivalent to | seán), (Jon (first name) | equivalent to | gianni (first name)), (Ioannis (name) | equivalent to | ifan (given name)), (Ioannis (name) | equivalent to | ion (given name)), (Ioannis (name) | equivalent to | Jean (first name)), (Ioannis (name) | equivalent to | johnny (first name))
Context1	The name Jon is equivalent to several other names, including Hans, Evan, jaan, and seán. Additionally, Jon is also synonymous with Gianni, which itself is equivalent to Jaan, Ivan, Juan, and Ioannis. Ioannis can further be linked to Ifan, Ion, Jean, and Johnny, showcasing a web of connections among these various names.
Context2	The name Gianni is equivalent to several other names, including Ivan, Jaan, Juan, and Ioannis. Additionally, it is noted that Ivan is different from Hans. Jon is another name that shares equivalencies, as it is equivalent to Evan, Jaan, Seán, and Gianni. Furthermore, the name Ioannis is equivalent to Ifan, Ion, Jean, and Johnny. This interconnected web of names shows the diverse ways in which names can correlate across different cultures and languages.

Table 13: Easy and Difficult Examples of MAGIC (1-conflict).