

Partial Perspectives: How LLMs Handle Logically Inconsistent Knowledge in Reasoning Tasks

Zichao Li¹, Ines Arous², Jackie C.K. Cheung^{1,3}

¹McGill University, MILA - Quebec AI Institute, ²York University, ³Canada CIFAR AI Chair
zichao.li@mila.quebec inesar@yorku.ca jackie.cheung@mcgill.ca

Abstract

Most natural language reasoning tasks in the research community assume consistent input knowledge. Nevertheless, real-world scenarios often involve inconsistent information, which might lead to divergent conclusions and are typically associated with varying levels of uncertainty. This raises a key research question: can large language models (LLMs) effectively handle uncertainty in their reasoning process to maximize knowledge consistency? In this paper, we propose a framework for evaluating reasoning over inconsistent knowledge. Our approach models uncertainty via weights of logical rules, leveraging Markov logic networks (MLN), which integrate probabilistic reasoning with first-order logic. This enables us to quantify inconsistencies in knowledge bases, and hence rigorously evaluate LLM reasoning. We introduce two tasks using this framework: 1) QA, which involves answering questions by integrating inconsistent knowledge; and 2) knowledge rectification, where we aim to rectify language models' acquired knowledge to improve consistency. We curate a dataset of 3,000 MLN-formatted knowledge bases to implement these tasks. We evaluate state-of-the-art LLMs on these tasks and highlight their limitations in uncertainty-aware reasoning over inconsistent logical knowledge.¹

1 Introduction

Natural language reasoning (Yu et al., 2024) forms a cornerstone of modern AI systems, enabling a wide range of applications such as question answering (Geva et al., 2021), clinical support (Gao et al., 2023), and digital agents (Su et al., 2024). Recent advancements in large language models (LLMs) (Kojima et al., 2022; Achiam et al., 2023) have led to impressive benchmark performance on tasks like commonsense reasoning (Bhagavatula et al., 2019) and logical reasoning (Parmar et al., 2024). Notably, most of these reasoning tasks assume that the input data is consistent and reliable. In practice, however, we frequently encounter inconsistent evidence, claims, or narratives, such as diverging health reports during COVID-19 (Nagler et al., 2020), varying interpretations of political and social issues (Pan et al., 2023), or evolving scientific discoveries that challenge prior findings (Kuhn, 1997).

Therefore, LLM-based applications need to process inconsistent information to derive well-reasoned conclusions. For example, an AI help desk responding to a network issue inquiry (see Figure 1) needs to evaluate competing answers based on different technical rules. The system should estimate the likelihood of each possible cause and guide the user toward a solution. Likewise, knowledge must adapt to new evidence that challenges existing beliefs (Konieczny & Pérez, 2005; Baral et al., 1992), such as revising taxonomies when genetic studies reveal unexpected relationships.

However, a rigorous definition of reasoning over inconsistent knowledge is nontrivial to pin down. A key challenge lies in defining a clear ground truth; that is, which competing conclusion the model should prioritize. The common NLP practice of relying on human annotations as a gold standard does not suffice here, since human judgments can

¹Code and data: <https://github.com/McGill-NLP/llm-logic-conflict>.

be influenced by cognitive biases, such as confirmation bias (Nickerson, 1998) and belief perseverance (Anderson et al., 1980) due to limited cognitive resource (Simon, 1955). Previous work (Xie et al., 2024; Qian et al., 2023) addresses this challenge using heuristics, such as prioritizing the most recent or user-provided information. However, heuristics are inherently rigid and fail to adapt to nuanced, uncertainty-aware scenarios common in real-world decision-making. For example, a rule may have lower priority than another *a priori*, yet its implications should carry more weight if numerous facts align with it.

To tackle this challenge, we draw on previous work in symbolic AI that handles uncertainty in logical reasoning. Specifically, we propose a novel framework based on the formulation of Markov logic networks (MLN) (Richardson & Domingos, 2006), in order to rigorously evaluate reasoning over inconsistent knowledge. MLN combines first-order logic rules with probabilistic reasoning such that we can assign weights to potentially inconsistent statements as a measure of uncertainty and precisely quantify a knowledge base’s consistency.

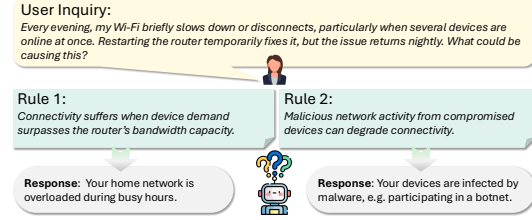


Figure 1: Contradicting conclusions derived using different rules.

Using this framework, we propose two tasks, consistency-enforced question-answering (QA) and knowledge rectification. The goal of QA is to answer questions through integrating potentially inconsistent knowledge. For example, in the scenario illustrated in Figure 1, each rule will be assigned a weight, and the system determines the most coherent response based on both the rule weights and the extent to which factual evidence aligns with them. In knowledge rectification, the model identifies and rectifies a specified number of facts for each knowledge base to maintain consistency. While the QA task aligns more closely with real-world use cases, the rectification task is more challenging, requiring reasoning across a high-dimensional space of possible resolutions and evaluating trade-offs for optimal consistency.

Both tasks are based on collections of *knowledge bases*, each of which provides factual and logical knowledge with varying levels of uncertainty. To implement these task setups, we curate a new dataset iKnow consisting of 3,000 knowledge bases, each of which contains factual statements and logical rules expressed in natural language. We conduct comprehensive evaluation of state-of-the-art open-sourced and proprietary LLMs. Our contributions are threefold: (1) We are the first to systematically study how LLMs reason with logically inconsistent knowledge using a Markov logic network formulation. (2) We propose two new tasks, consistency-enforced QA and knowledge rectification, and release accompanying datasets. (3) We show in our experiments that current LLMs often fail to comprehensively resolve inconsistent knowledge. Instead, they tend to selectively pick a subset of the knowledge base, drawing conclusions based solely on that subset while ignoring the rest.

2 Related Work

2.1 Knowledge Conflicts

There has been a growing body of work studying knowledge inconsistencies in LLMs. Xu et al. (2024) taxonomize and categorize these inconsistencies into context-memory conflicts (Longpre et al., 2021), inter-context conflicts (Xie et al., 2024; Qian et al., 2023), and intra-memory conflicts (Qi et al.). They highlight conflicts arising from different sources of information, including parametric knowledge and external information in prompts. These studies reveal LLMs’ susceptibility to confirmation bias; for instance, Xie et al. (2024) report that LLMs tend to favor external information that aligns with their parametric knowledge. Another closely related area of research is fact-checking, which aims to identify discrepancies between claims and supporting evidence drawn from textual (Thorne et al., 2018; Wadden et al., 2020) or tabular sources (Chen et al., 2019).

Previous work focused primarily on factual knowledge conflicts, such as the number of Olympic medals a country has won. Additionally, these studies often adopt a simplistic, binary view of conflict, which usually stems from misinformation or outdated facts, overlooking the inconsistencies caused by knowledge uncertainty. By contrast, our proposed framework extends beyond the binary case, offering a more general approach to reasoning over inconsistent knowledge, including contradictions and factual-logical discrepancies, by modeling varying degrees of uncertainty. In other words, it requires LLMs to weigh competing hypotheses in uncertain logical contexts.

2.2 Logical Reasoning with Natural Language

There has been increasing interest in using LLMs for logical reasoning. Natural language provides a flexible interface for injecting rule-based knowledge to guide or control neural models, which is a fundamental challenge when using neural AI systems. Recent studies have explored various types of logical reasoning in neural language models, such as abductive reasoning (Bhagavatula et al., 2019), defeasible inference (Rudinger et al., 2020), deductive reasoning (Clark et al., 2021), and theorem proving (Tafjord et al., 2021). However, existing research often assumes the provided logical knowledge is consistent, overlooking reasoning in the presence of contradictions. More recently, there has been growing interest on resolving logical conflicts within LLM outputs (Jung et al., 2022; Mitchell et al., 2022; Kassner et al., 20223). Conversely, we examine how LLMs handle inconsistent logical knowledge within input.

2.3 Reasoning with Inconsistent Logical Knowledge

Logical reasoning with uncertainty has been a long-standing problem in machine learning and AI. Richardson & Domingos (2006) propose Markov logic networks to address ambiguities caused by inconsistent logical knowledge with uncertainty. Since then, MLNs have been applied across various domains with applications such as KB-based question answering (Poon & Domingos, 2009), opinion extraction (de la Vega & Ng, 2016), and video action reasoning (Jin et al., 2022). In this work, we leverage MLNs to establish a framework for evaluating LLMs in reasoning over inconsistent knowledge.

3 Evaluating Reasoning over Logically Inconsistent Knowledge

We start by specifying guidelines to help us define what it means for models to reason over inconsistent knowledge. Concretely, we ground our approach in three key principles: (1) *maximal internal consistency* within our knowledge of the world (Harmon-Jones & Mills, 1999; Peppas, 2008); (2) integrating *uncertainty* into the reasoning process (as opposed to binary true-false beliefs) (Dempster, 1968); and (3) *prioritizing newly acquired knowledge* as it becomes available (Piaget, 1977; Delgrande et al., 2004). Driven by these principles, we formally define reasoning over inconsistent knowledge, then introduce two tasks within this formulation.

3.1 Formulation

We first define a first-order logic knowledge base $K = \{R, F\}$, where $R = \{r_1, r_2, \dots, r_n\}$ is a set of first order logical rules expressed in natural language, and $F = \{f_1, f_2, \dots, f_m\}$ a set of atomic formulas, also expressed in natural language, and related to the rules in R . For simplicity, we refer to them as factual statements below.

Example 1. The set R is composed of the following rules:

- $r_1 = \text{If a [Person] works at [Company], and [Company] is located in [City], then the [Person] is a resident of [City].}$
- $r_2 = \text{If [Person] was born in [City], then they are the resident of [City].}$

The set F is composed of the following facts related to R :

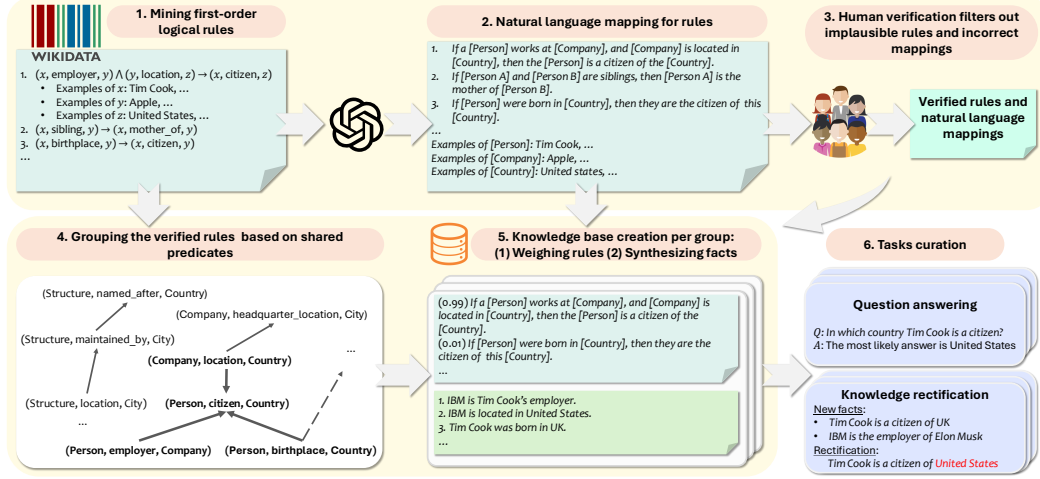


Figure 2: The pipeline of the iKnow curation. Knowledge base creation: (1) We first mine a large set of first-order logic rules candidates from Wikidata (methods detailed in Section 4.2.1). (2 & 3) Then, we generate their natural language mappings and infer the variable types; Later, human annotators filter out the implausible rules and incorrect mappings (Section 4.2.1). (4 & 5) We group verified rules based on shared relations, then create a KB per group of rules, synthesizing facts for each relation in the rules and assign weights on rules (Section 4.2.3). Task construction: we curate two task instances for each KB. We use a MLN engine to derive ground-truth answers and rectifications (Section 4.3).

- *Alice works at SmartAI; SmartAI is located in Boston; Alice was born in Chicago.*

Each rule represents an *If-Then* logical connection between *premises* and *implications*. For instance, r_2 corresponds to the symbolic form $(\text{Person}, \text{birthplace}, \text{City}) \rightarrow (\text{Person}, \text{residency}, \text{City})$, where *Person* and *City* are *variables* that must be instantiated with entities of their respective types, such as *Alice* and *Chicago*. The relations *birthplace* and *residency* define relationships between these variables. A fact is a grounding of a relation where variables are instantiated, e.g. $(\text{Alice}, \text{birthplace}, \text{Chicago})$.

Given a query q , the model infers a set of consequent facts \hat{F} with uncertainty, e.g. *Alice lives in the city* $_$, based on K , maximizing consistency in a knowledge base. We denote $S(K \cup \hat{F})$ as a measure of the internal consistency within a given knowledge base. As discussed above, $S(\cdot)$ should incorporate uncertainty of knowledge and hence provide a nuanced quantification of consistency.

In this work, we use the formulation of Markov logic networks (MLNs) that allows users to assign priorities in terms of weights w_i to each logical rule r_i , enabling the integration of uncertainty directly into first-order logic rules. This flexibility makes MLNs broadly applicable across domains, including AI (Poon & Domingos, 2009; de la Vega & Ng, 2016; Jin et al., 2022), biology (Sakhanenko & Galas, 2012), and IoT (Ala et al., 2024). Formally, a Markov logic network L is a set of weighted rules (r_i, w_i) . Together with a set of entities E , it defines a Markov network $M_{L,E}$ with one node per possible grounding of each relation in L , e.g. $(\text{Alice}, \text{employer}, \text{SmartAI})$. The probability of a particular set of facts F is then:

$$P_{M_{L,E}}(F) = \frac{1}{Z} \exp \left(\sum_i w_i c_i(x) \right), \quad (1)$$

where $c_i(x)$ is the number of true groundings of rule r_i in facts F , and Z is a normalization constant. Since computing Z is intractable in practice, we define the consistency function S

as an unnormalized score function (Richardson & Domingos, 2006):

$$S_{M_{L,E}}(F) = \exp \left(\sum_i w_i c_i(x) \right). \quad (2)$$

Throughout this paper, *uncertainty* refers to the weights assigned to each rule (i.e., w_i), while *inconsistency* describes the logical state arises from applying different rules. Based on these definitions and the MLN formulation, we introduce two tasks, question-answering (QA) and knowledge rectification.

3.2 Task 1: Consistency-enforced QA

In the QA task, the model predicts an answer a to the query q about an entity, e.g. *In which city does Alice live?* conditioned on a knowledge base (KB) $K = \{M_{L,E}, F\}$, inferring an unknown fact $\hat{f} = (q, a)$. The goal of the model is to maximize the consistency of \hat{f} with respect to the KB, combining the influence of both weighted rules and facts relevant to q :

$$\arg \max_a S_{M_{L,E}}(F \cup \hat{f}). \quad (3)$$

3.3 Task 2: Knowledge Rectification

The rectification task reflects scenarios where a KB adapts to new information while managing potential inconsistencies. Ideally, the update should adhere to the principles introduced in Sec. 3, where we prioritize maintaining the consistency of the KB while also balancing the retention of prior knowledge, and integrating new facts (Dalal, 1988). Specifically, the goal of this task is to rectify up to a fixed number b of statements in a new fact set F' given the original K , such that the resulting \hat{F}' maximizes the internal consistency of the updated knowledge base:

$$\arg \max_{\hat{F}'} S_{M_{L,E}}(F \cup \hat{F}') \quad \text{s.t.} \quad |\hat{F}' \triangle F'| \leq b, \quad (4)$$

where \triangle denotes the symmetric difference operator.

Logical rules and assigned weights	Rule 1: <i>If a [Person] works at [Company], and [Company] is located in [Country], then the [Person] is a citizen of the [Country].; Weight: 1.</i> Rule 2: <i>If [Person] were born in [Country], then they are the citizen of this [Country].; Weight: 9999.</i>
Known Facts	1. IBM is Tim Cook's employer. 2. IBM is located in United States. 3. Tim Cook was born in UK.
Query Expected Answer	<i>In which country Tim Cook is a citizen?</i> UK
New Facts Expected Rectification	{ Tim Cook is a citizen of United States; IBM is the employer of Elon Musk. } { Tim Cook is a citizen of UK . }

Table 1: A simplified example from iKnow includes a knowledge base and two tasks, QA and knowledge rectification, curated from it. Please refer to Table 4 for a real example.

4 iKnow: A Test Suite for Reasoning over Inconsistent Knowledge

To implement the proposed task formulation, we develop iKnow, a test suite for evaluating the ability of LLMs to reason with possibly inconsistent logical **Knowledge**.

4.1 Overview

Our test suite iKnow is developed in two main stages: knowledge base (KB) construction and task instance curation based on each KB. Each KB is composed of two elements: *logical*

rules and grounding facts. Table 1 illustrates these two elements with a simplified example, and Figure 2 provides an overview for each construction step. In Section 4.2, we explain each step for constructing both elements in detail. First, we construct the *logical rules* in Section 4.2.1, where we mine first-order logic rules and map them to their natural language form (e.g. the first row in Table 1). Then, we cluster them based on common relations and subsequently synthesize the *grounding facts* in Section 4.2.2 (e.g. the second row in Table 1).

4.2 Knowledge Base curation

4.2.1 Constructing Logical Rules

We aim to collect diverse and plausible logical rules that to some extent capture dependencies among real-world facts. We construct a comprehensive set of first-order logic rules along with their expressions in natural language as seen in Table 1. This process involves two main steps: symbolic rule mining and natural language mapping:

Symbolic Rule Mining We mine logical rules, *If Premises then Implications*, from Wikidata (Vrandečić & Krötzsch, 2014) by enumerating all combinations of relations that fit the following first-order logic templates: **Subrelation**: $(e_x, r_1, e_y) \rightarrow (e_x, r_2, e_y)$; **Symmetric**: $(e_x, r_1, e_y) \rightarrow (e_y, r_1, e_x)$; **Existential**: $(e_x, r_1, e_y) \rightarrow \exists z(e_x, r_2, e_z) \wedge (e_z, r_3, e_y)$; **Negation**: $(e_x, r_1, e_y) \rightarrow \neg(e_x, r_2, e_y)$ and $(e_x, r_1, e_y) \rightarrow \neg(e_y, r_2, e_x)$; **Combination**: $(e_x, r_1, e_y) \wedge (e_y, r_2, e_z) \rightarrow (e_x, r_3, e_z)$, where e_x, e_y and e_z are variables and r denotes the relation. We keep only the rules supported by a sufficient number of facts in Wikidata. Appendix B.1 provides more details. As demonstrated by Qu & Tang (2019), this simple mining method is sufficient to automatically extract large-scale rule candidates that are useful to derive unknown facts in Wikidata. In addition, we collect entities in supporting facts, such as *Tim Cook* and *United States* shown in (1) of Figure 2, as entities that can be used to instantiate the variables.

Natural Language Mapping This step aims to generate natural language representations of first-order logic rules, which serve as input in the proposed tasks. We leverage GPT-4o (Hurst et al., 2024) to generate textual mappings by providing it with rules and potential instantiations of their variables. The model infers variable types, such as *[Company]* and *[person]*, and produces structured natural language expressions for the rules. For example, $(e_x, \text{employer}, e_y) \wedge (e_y, \text{country}, e_z) \rightarrow (e_x, \text{citizen}, e_z)$ is transformed to Rule 1 in Table 1.

To ensure high-quality logical rules, we filter out implausible rules using crowdsourcing. Each rule, presented in natural language, is evaluated by two annotators who classify it into one of five potential categories: *Unrelated*, *Contradictory*, *Less Likely*, *More Likely*, or *Necessarily True*. Rather than enforcing strict logical entailment, we assess whether LLMs can reason over user-provided rules and their assigned weights. However, we aim for the rules in iKnow to reflect plausible real-world scenarios. Therefore, we discard any rule labeled *Contradictory* by at least one annotator (e.g., *If [City] is the capital of [Country], then [City] is not located in [Country]*). To maintain high annotation quality, we implement a qualification test for annotators. This test includes rules with known ground-truth labels and edge cases designed to assess logical reasoning skills. Only candidates who meet our high performance standards are allowed to participate in the main task. Appendix B.2 provides more details of crowdsourcing setup. Additionally, we integrate logical rules and their natural language expressions from DepEdit (Li et al., 2023), which provides post-verified rules to further enlarge the dataset. Overall, this process results in a set of approximately 2,200 logical rules in first-order logical form together with their natural language mapping.

4.2.2 Grouping and Weighting Rules

After we collect a large set of logical rules and their natural language equivalents, we group them to create knowledge bases based on shared relations. For instance, Rules 1 and 2 in Table 1 share the same implications but different premises. To do so, we begin with a single rule for creating each knowledge base and perform a depth-first search to retrieve additional

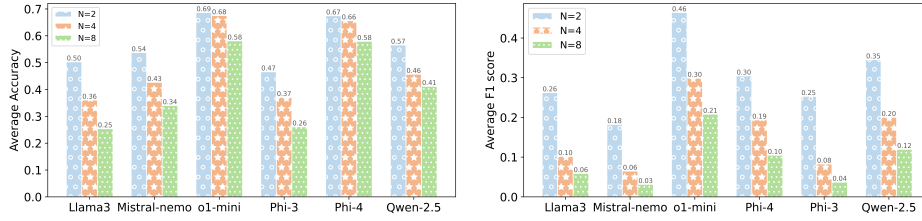


Figure 3: The effect of the number of facts per relation on the LLMs’ performance on QA and rectification tasks.

rules consisting of overlapping relations. In total, we extract approximately 3000 groups of rules, each of them contain five rules in average.

In addition, we assign a weight for each rule in the KB. In practice, these weights can be determined using data-driven methods (Huynh & Mooney, 2008; Davis & Domingos, 2010; Qu & Tang, 2019) or by incorporating domain knowledge from human experts (Tran & Davis, 2008; Singla & Mooney, 2011; Khot et al., 2015). Since our focus is on evaluating whether LLMs can incorporate these weights, rather than obtaining accurate weights, we adopt three artificial schemes to synthesize weights: (1) equal, where all the rules share the same weight, (2) polarized, where some randomly selected rules have extreme high weights such that they should be always respected, (3) random, where the weights are sampled from a Normal distribution $\mathcal{N}(10, 1)$.

4.2.3 Synthesizing Facts

We generate factual statements, e.g. the second row in Table 1, to ground each relation in the rules. To do so, we first transform triplets of variables and relations, e.g. $([Person], \text{employer}, [Company])$ into statement templates, e.g. $[Company]$ is $[Person]$ ’s employer, using GPT-4o, then randomly sample entities to instantiate the variables, e.g. instantiate $[Person]$ with *Tim Cook*. This approach allows us to flexibly reuse a fixed set of five statement templates per relation, enabling scalable fact synthesis.

4.3 Task construction using knowledge bases

Next, we construct a task instance for QA and knowledge rectification, respectively for each KB. For the QA task, we include the synthesized facts of relations in premises in input and queries require the model to derive implications. For instance, the query in Table 1 requires inferring the implication in Rules 1 and 2. Queries are generated similarly to factual statements: we convert triplets of entity types and relations into query templates, e.g. *In which country [person] is a citizen*, then substitute the placeholder with specific entities, *Tim Cook*. In the end, we derive the most likely answer (*United States* in Table 1). For the knowledge rectification task, we first derive implications from all premises, then corrupt $m \sim \text{Uniform}\{1, 2\}$ of them by substituting entities with others of the same type ($UK \Rightarrow \text{United states}$ in Table 1).

To determine ground-truth answers or rectifications, we use Tuffy (Niu et al., 2011), a Markov logic network engine that processes symbolic knowledge, including rules with their associated weights and facts, to compute the likelihood of unknown facts.

5 Experiments

We experiment with iKnow to investigate whether LLMs are able to successfully integrate uncertainty into reasoning over logically inconsistent knowledge.

5.1 Experiment setup

Base models We experiment with the most widely-used LLMs that excel in reasoning benchmarks, including Llama3_{8B} (Dubey et al., 2024), Mistral-nemo (AI, 2025), Phi-3_{14B} (Abdin et al., 2024a), Qwen-2.5_{14B} (Yang et al., 2024), Phi-4_{14B} (Abdin et al., 2024b) and o1-mini (OpenAI, 2024). We adopt both one-shot in-context learning (Achiam et al., 2023) and chain-of-thought reasoning (Wei et al., 2022) as the inference method for open-sourced LLMs and CoT only for o1-mini.

Dataset variants To comprehensively evaluate the ability of LLMs to resolve knowledge inconsistencies, we define three key aspects of our dataset: (1) The number of facts per relation, $N \in \{2, 4, 8\}$, which determines the total number of facts included in the knowledge base and controls its size. (2) Rule weighting schema that governs the relative priority of each rule. As detailed in Section 4.2.3, we adopt three variants: *Equal*, *Polarized*, *Random*. (3) As for the existence of inconsistency among logical knowledge, we construct two variants: *conflict*, where the derived implications can be conflicting (e.g. Table 1), and *consist* without inconsistencies among logical rules.

Metrics Since the ground-truth answers or rectifications are short and we found that LLMs are able to generate output in the correct format, we adopt simple lexicon-matching metrics. In particular, we adopt the exact-match (EM) score for QA. As for knowledge rectification, we measure precision, recall, and F1-score between the predicted \hat{F}' and ground-truth rectifications. Appendix D provides more experiment details.

5.2 Performance of LLMs on iKnow

Figure 3 provides an overview of LLM performance on iKnow, grouped by the number of facts per relation. As shown, the top-performing models are o1-mini and Phi-4, consistent with their superior results on reasoning benchmarks compared to other models. Notably, all LLMs perform reasonably well when only a few facts are involved, achieving an average EM score above 0.50. However, performance drops sharply (20%-50%) as the number of facts increases. When confronted with more complex reasoning in knowledge rectification, which requires handling inconsistencies across multiple implications, performance declines significantly. Even the best-performing model, o1-mini, only achieves an F1 score of 0.46 with $N = 2$. Meanwhile, models like Llama3, Mistral, and Phi-3 exhibit near-zero performance when $N = 8$. Figure 4 demonstrates the impact of rule weighting schema on LLMs’ performance on QA. While *Polarized* is expected to be the simplest schema, where the LLM can easily follow the rule with the dominant weight, we found no significant performance differences across schemas. A similar trend appears in the rectification task Figure 8. As we will show in more in-depth analysis later, this is mainly because LLMs struggle to integrate knowledge uncertainty into the reasoning process.

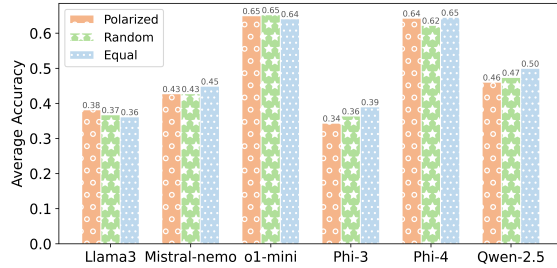


Figure 4: The effect of the rule weighting schema on the LLMs’ performance on QA.

5.3 Analysis and Findings

Next, we conduct a fine-grained analysis to understand why LLMs perform poorly on iKnow and present four key findings below.

Finding 1. LLMs are unable to emulate MLNs in reasoning over inconsistent logical knowledge.

To investigate the direct effect of inconsistent logical knowledge on LLMs’ reasoning, Figure 6 compares their performance on the *conflict* and *consistent* splits for both tasks. The

results reveal a significant performance drop (by 20% on average) when conflicting implications arise from different logical rules. In particular, the F1-score of o1-mini decreases by 33% on *conflict* split. To further investigate whether LLMs genuinely comprehend the meaning of rule weights, we conducted a targeted experiment to confirm LLMs recognize that higher weights correspond to higher priorities. Appendix E provides more details.

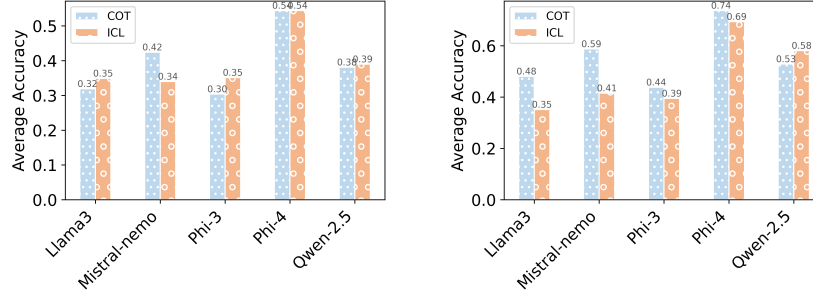


Figure 5: The effect of decoding methods on the QA accuracy.

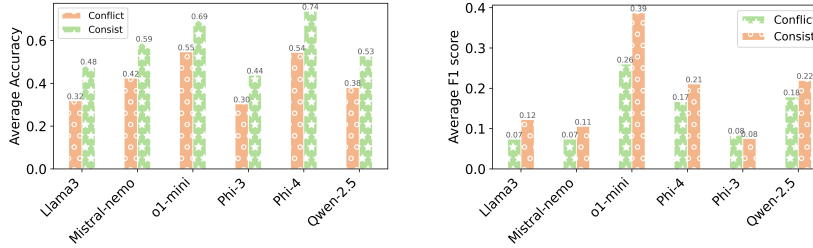


Figure 6: The effect of conflicting implications on LLMs’ performance of QA (left) and knowledge rectification (right).

Input – rules	<p>Rule 1: If an [artwork] is part of a [collection] and this [collection] is named after a [person], then this [person] is proposed as the creator of the [artwork]. <i>Weight: 10.</i></p> <p>Rule 2: If [artwork A] is described by [artwork B] and [artwork B] is created by [person], then [artwork A] is also created by that [person]. <i>Weight: 10.</i></p> <p>Rule 3: If an [artwork] is part of a [collection] and that [collection] was created by a [person], then that [artwork] was also created by that [person]. <i>Weight: 10.</i></p> <p>...</p>
Input – known facts	(1) Wassily Kandinsky is credited as the maker of Tate Modern; (2) Nighthawks is described by Canon of Medicine; (3) Jacqueline Picasso (Picasso)’s notable work includes Canon of Medicine; (4) Tate Modern is named after Wassily Kandinsky; (5) Bosch Collection is named after Jacqueline Picasso (Picasso); (6) Nighthawks belongs to the Tate Modern archival collection. ...
Input – queries	Who has been identified as the original creator for the Nighthawks?
LLM output	... 1. We know that “Nighthawks is described by Canon of Medicine”. According to rule 2, ... we can infer that the creator of “Canon of Medicine” is also the creator of “Nighthawks”. 2. We are told that “Jacqueline Picasso (Picasso)’s notable work includes Canon of Medicine”. Therefore, according to our inference from step 1, Jacqueline Picasso (Picasso) is also the creator of “Nighthawks”. Final answer: Jacqueline Picasso (Picasso)
Ground-truth	Wassily Kandinsky

Table 2: An example output from Mistral-nemo illustrates an error in reasoning due to a subset of logical rules being omitted. The missing rules are highlighted in red. Table 6 demonstrates a same type of error from o1-mini.

Finding 2. Advanced decoding methods, such as chain-of-thought reasoning, do not enhance LLMs’ ability to handle inconsistent logical knowledge.

Next, we test whether more advanced decoding methods help LLMs reach more coherent conclusions. Interestingly, Figure 5 shows that CoT does not improve accuracy and, in some

cases, even worsens it on the QA task, except for Mistral-Nemo. However, CoT has a more positive effect on the *consistent* variant. As further demonstrated by Figure 9, CoT does not improve LLMs’ performance on the rectification task neither. As we will show later in *Finding 4*, this is mainly because CoT sometimes activates their prior knowledge that distracts the reasoning and leads it off track.

Finding 3. *LLMs tend to apply only a subset of relevant rules during the reasoning process.*

To better understand why LLMs struggle with inconsistent logical knowledge, we conduct an in-depth error analysis. We examine reasoning chains generated by Mistral-Nemo that lead to incorrect final answers and categorize the mistakes into two main types: *omit relevant rules* and *hallucination*. Among 50 incorrect reasoning artifacts, 43 fall into the *omit relevant rules*, where the model considers only a subset of applicable rules, while 7 are due to *hallucination*, where the model either fails to recall the correct fact from the input or relies on its parametric knowledge instead of the given premises to derive the most plausible conclusion. Table 2 illustrates an example of omitting relevant rules, and Table 7 provides additional examples of both error types. To further validate this finding, we conduct a quantitative analysis to measure how often LLMs reason using only a subset of applicable rules. Specifically, we randomly sample 50 incorrect predictions from Mistral, o1-mini, and Llama3, then determine whether all relevant rules were considered in the derivation. Table 3 presents the results, suggesting that more than half of the erroneous are due to missing applicable rules into reasoning process.

Finding 4. *LLMs tend to insist on their parametric knowledge during derivation.*

Notably, the performance of LLMs are generally poor, even on the consistent variant of iKnow, which contains no contradictory implications. Our qualitative analysis reveals that LLMs tend to favor commonsense knowledge from pre-training over implications derived from given rules and facts when conflicts arise. Table 8 illustrates this with a concrete example. This finding aligns with previous research (Longpre et al., 2021) that reports that LLMs prefer their pre-trained factual knowledge over the conflicting one in the input. Additionally, LLMs may “activate” their parametric knowledge as premises, influencing their reasoning. For instance, as shown in Table 9, the model prioritizes its prior knowledge, *Mona Lisa was created by Leonardo da Vinci*, over the correct reasoning chains.

Moreover, Table 10 demonstrates that models rely on inherent logical assumptions, such as *Two people speaking the same language are likely to work for the same organization*, instead of the applicable logical rules provided in the prompt.

Models	Omit relevant rules (%)
Mistral-nemo	86.00
Llama3 _{8B}	76.00
OpenAI-o1-mini	56.00

Table 3: Percentage of reasoning with incomplete set of applicable rules.

6 Conclusion

This study evaluates LLMs’ reasoning over inconsistent logical knowledge using a Markov logic network formulation. We introduce two new tasks: consistency-enforced QA and knowledge rectification, and curated an accompanying dataset iKnow. Experiments reveal that LLMs fail to sufficiently incorporate the uncertainty in their knowledge to infer in a way that maximizes knowledge consistency. The primary issue is their inability to incorporate all relevant rules. Other limitations include hallucinations and a tendency to rely on prior factual and logical knowledge. Our work can be extended to incorporate more complex reasoning tasks. We discuss the limitations and potential future directions in Appendix A.

7 Ethics Statement

Our work aims to advance LLMs’ reasoning over inconsistent knowledge. However, models capable of reasoning about uncertainty may introduce new or greater harms compared

to the current generation of LLMs. For instance, imperfect models might provide inaccurate uncertainty assessments, potentially leading to harmful outcomes. With the rapid development of LLMs, it is unsurprising that their performance on our dataset iKnow may significantly improve in the near future. However, this progress may not decrease the risk of harm during development and deployment, especially in complex or specific scenarios, since these potential harms are not captured by our dataset or evaluation.

Moreover, our dataset, curated from Wikidata, covers only a limited domain of relations and entities and inherently reflects biases from the original source. The collected logical rules are intended solely to evaluate LLMs' reasoning with provided knowledge and uncertainty. They do not represent strict logical entailment and should not be used for decision-making purposes.

While future work must remain mindful of these potential harms, we believe our evaluation framework and dataset can support ongoing research in this field. This is crucial for developing responsible AI systems, especially as LLMs are increasingly integrated into automated decision-making.

Acknowledgment

We thank Cesare Spinoso-Di Piano for his invaluable feedback on our manuscript. We also thank the Area Chair and the reviewers for their thoughtful comments and suggestions.

References

- Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024a.
- Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024b.
- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- Mistral AI. Mistral nemo, Jan 2025. URL <https://mistral.ai/news/mistral-nemo/>.
- Ali Ala, Amir Hossein Sadeghi, Muhammet Deveci, and Dragan Pamucar. Improving smart deals system to secure human-centric consumer applications: Internet of things and markov logic network approaches. *Electronic Commerce Research*, 24(2):771–797, 2024.
- Craig A Anderson, Mark R Lepper, and Lee Ross. Perseverance of social theories: The role of explanation in the persistence of discredited information. *Journal of personality and social psychology*, 39(6):1037, 1980.
- Chitta Baral, Sarit Kraus, Jack Minker, and Venkataram Subram Subrahmanian. Combining knowledge bases consisting of first-order theories. *Computational intelligence*, 8(1):45–71, 1992.
- Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. Abductive commonsense reasoning. *arXiv preprint arXiv:1908.05739*, 2019.
- Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyu Zhou, and William Yang Wang. Tabfact: A large-scale dataset for table-based fact verification. *arXiv preprint arXiv:1909.02164*, 2019.

- Peter Clark, Oyvind Tafjord, and Kyle Richardson. Transformers as soft reasoners over language. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pp. 3882–3890, 2021.
- Mukesh Dalal. Investigations into a theory of knowledge base revision: preliminary report. In *Proceedings of the Seventh AAAI National Conference on Artificial Intelligence*, pp. 475–479, 1988.
- Jesse Davis and Pedro Domingos. Bottom-up learning of markov network structure. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 271–280. Omnipress, 2010.
- Luis Gerardo Mojica de la Vega and Vincent Ng. Markov logic networks for text mining: A qualitative and empirical comparison with integer linear programming. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pp. 4388–4395, 2016.
- James Delgrande, Torsten Schaub, Hans Tompits, and Kewen Wang. A classification and survey of preference handling approaches in nonmonotonic reasoning. *Computational Intelligence*, 20(2):308–334, 2004.
- Arthur P Dempster. A generalization of bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2):205–232, 1968.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- Yanjun Gao, Dmitriy Dligach, Timothy Miller, John Caskey, Brihat Sharma, Matthew M Churpek, and Majid Afshar. Dr. bench: Diagnostic reasoning benchmark for clinical natural language processing. *Journal of biomedical informatics*, 138:104286, 2023.
- Mor Geva, Daniel Khoshdel, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021.
- Eddie Harmon-Jones and Judson Mills. Cognitive dissonance. *Progress on a pivotal theory in social psychology*. Washington, DC: American Psychological Association, 1999.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- Tuyen N Huynh and Raymond J Mooney. Discriminative structure and parameter learning for markov logic networks. In *Proceedings of the 25th international conference on Machine learning*, pp. 416–423, 2008.
- Yang Jin, Linchao Zhu, and Yadong Mu. Complex video action reasoning via learnable markov logic network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3242–3251, 2022.
- Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. Maieutic prompting: Logically consistent reasoning with recursive explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 1266–1279, 2022.
- Nora Kassner, Oyvind Tafjord, Ashish Sabharwal, Kyle Richardson, Hinrich Schuetze, and Peter Clark. Language models with rationality. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.
- Tushar Khot, Niranjan Balasubramanian, Eric Gribkoff, Ashish Sabharwal, Peter Clark, and Oren Etzioni. Exploring markov logic networks for question answering. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pp. 685–694, 2015.

- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.
- Sébastien Konieczny and Ramón Pino Pérez. Propositional belief base merging or how to merge beliefs/goals coming from several sources and some links with social choice theory. *European Journal of Operational Research*, 160(3):785–802, 2005.
- Thomas S Kuhn. *The structure of scientific revolutions*, volume 962. University of Chicago press Chicago, 1997.
- Zichao Li, Ines Arous, Siva Reddy, and Jackie Chi Kit Cheung. Evaluating dependencies in fact editing for language models: Specificity and implication awareness. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7623–7636, 2023.
- Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.
- Eric Mitchell, Joseph Noh, Siyan Li, Will Armstrong, Ananth Agarwal, Patrick Liu, Chelsea Finn, and Christopher D Manning. Enhancing self-consistency and performance of pre-trained language models through natural language inference. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 1754–1768, 2022.
- Rebekah H Nagler, Rachel I Vogel, Sarah E Gollust, Alexander J Rothman, Erika Franklin Fowler, and Marco C Yzer. Public perceptions of conflicting information surrounding covid-19: Results from a nationally representative survey of us adults. *PloS one*, 15(10): e0240776, 2020.
- Raymond S Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220, 1998.
- Feng Niu, Christopher Ré, AnHai Doan, and Jude Shavlik. Tuffy: Scaling up statistical inference in markov logic networks using an rdbms. *Proceedings of the VLDB Endowment*, 4(6), 2011.
- OpenAI. Learning to reason with llms, Sep 2024. URL <https://openai.com/index/learning-to-reason-with-llms>.
- Jinsheng Pan, Weihong Qi, Zichen Wang, Hanjia Lyu, and Jiebo Luo. Bias or diversity? unraveling fine-grained thematic discrepancy in us news headlines. *arXiv preprint arXiv:2303.15708*, 2023.
- Mihir Parmar, Nisarg Patel, Neeraj Varshney, Mutsumi Nakamura, Man Luo, Santosh Mashetty, Arindam Mitra, and Chitta Baral. Logicbench: Towards systematic evaluation of logical reasoning ability of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13679–13707, 2024.
- Pavlos Peppas. Belief revision. *Foundations of Artificial Intelligence*, 3:317–359, 2008.
- Jean Piaget. *The development of thought: Equilibration of cognitive structures*. (Trans A. Rosin). Viking, 1977.
- Hoifung Poon and Pedro Domingos. Unsupervised semantic parsing. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pp. 1–10, 2009.
- Jirui Qi, Raquel Fernández, and Arianna Bisazza. Cross-lingual consistency of factual knowledge in multilingual language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.
- Cheng Qian, Xinran Zhao, and Sherry Tongshuang Wu. "merge conflicts!" exploring the impacts of external distractors to parametric knowledge graphs. *arXiv preprint arXiv:2309.08594*, 2023.

- Meng Qu and Jian Tang. Probabilistic logic neural networks for reasoning. *Advances in neural information processing systems*, 32, 2019.
- Matthew Richardson and Pedro Domingos. Markov logic networks. *Machine learning*, 62: 107–136, 2006.
- Rachel Rudinger, Vered Shwartz, Jena D Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A Smith, and Yejin Choi. Thinking like a skeptic: Defeasible inference in natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4661–4675, 2020.
- Nikita A Sakhanenko and David J Galas. Probabilistic logic methods and some applications to biology and medicine. *Journal of Computational Biology*, 19(3):316–336, 2012.
- Herbert A Simon. A behavioral model of rational choice. *The quarterly journal of economics*, pp. 99–118, 1955.
- Parag Singla and Raymond Mooney. Abductive markov logic for plan recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, pp. 1069–1075, 2011.
- Yu Su, Diyi Yang, Shunyu Yao, and Tao Yu. Language agents: Foundations, prospects, and risks. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pp. 17–24, 2024.
- Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. Proofwriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3621–3634, 2021.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 809–819, 2018.
- Son D Tran and Larry S Davis. Event modeling and recognition using markov logic networks. In *Computer Vision—ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12–18, 2008, Proceedings, Part II 10*, pp. 610–623. Springer, 2008.
- Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.
- David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. *arXiv preprint arXiv:2004.14974*, 2020.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7–11, 2024*. OpenReview.net, 2024. URL <https://openreview.net/forum?id=auKAUJZM06>.
- Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for LLMs: A survey. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8541–8565, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.486. URL <https://aclanthology.org/2024.emnlp-main.486/>.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.
- Fei Yu, Hongbo Zhang, Prayag Tiwari, and Benyou Wang. Natural language reasoning, a survey. *ACM Computing Surveys*, 56(12):1–39, 2024.

A Limitation

Evaluation framework Our work proposes a framework to evaluate LLMs’ reasoning over inconsistent and uncertain knowledge. As a first step toward this goal, our framework makes certain assumptions and has constraints that may not hold in real-world applications. First, we assume the weights are provided by users. While this allows the injection of domain knowledge, as shown in prior work (Tran & Davis, 2008; Singla & Mooney, 2011; Khot et al., 2015), LLMs often need to estimate these weights autonomously in practical settings, where they must assess information sources or supporting facts. Existing machine learning methods can be adapted for this purpose (Huynh & Mooney, 2008; Davis & Domingos, 2010; Qu & Tang, 2019), which we leave for future work. Second, we assume all knowledge required for inference is present in the provided knowledge base. In practice, this can be achieved through prompting or retrieval augmentation. However, it is possible that the provided knowledge is incomplete, requiring the model to combine parametric and external knowledge for sufficient reasoning. Third, the rules are constrained by their first-order nature, meaning they cannot express quantification over sets or relations, nor can they handle complex arithmetic operations.

Dataset curation We collect logical rules from Wikidata in the domain of commonsense knowledge using simple heuristics. As a result, these rules are less complex than those encountered by LLM-based systems in practice, such as the domains of law, medicine, and mathematics. Following recent practices, parts of our dataset are synthesized using GPT-4o, which may be biased and limit the diversity compared to real-world user queries. Lastly, our dataset is exclusively in English.

Experiments and findings Although we have conducted extensive experiments with six LLMs, it is important to note that our findings are specific to the investigated task and these particular LLMs, and should not be assumed universally applicable to other tasks or models. While our framework is general, we leave the exploration of more task varieties and additional LLMs for future work. Our experiments reveal novel and important insights, showing that LLMs tend to focus on a subset of knowledge when reasoning over inconsistent information. However, we do not yet understand why they prioritize certain rules, and we leave this deeper investigation using advanced interpretability tools to future effort. Moreover, our study does not evaluate human performance as a point of reference. It is mainly because our task setting goes beyond the paradigm of conventional NLP, which assumes that NLP systems should aim to replicate human understanding and use of language. Instead, our tasks involves a type of reasoning that people would find difficult, since our scenario features inherently inconsistent input knowledge, where humans are subject to cognitive biases, such as confirmation bias and belief perseverance, that result in flawed judgments. Also, the task needs precise computation of probabilities of competing hypotheses, which would be challenging for standard crowdsourcing workers. However, these issues which cause difficulties for human annotations are also precisely the reasons why our task is interesting and relevant, as LLMs that perform well on this task would have high practical utility in augmenting what people could do. That said, we believe that it would be interesting for future work to explore how human reasoning handles these logical inconsistencies compared to LLMs.

B Details of iKnow Construction

B.1 Logical Rule Mining

We build upon the brute-force search method from Qu & Tang (2019) to mine first-order logic rules. Specifically, we enumerate all candidate rules by combining relations from Wikidata, following the rule templates introduced in Section 4.2.1. We then select rules with sufficient factual grounding, measured by

$$\frac{\# \text{Implication} = \text{True} \wedge \text{Premise} = \text{True}}{\# \text{Premise} = \text{True}}$$

, with a threshold of 0.6. However, this method only applies to rules with universal quantification. To mine rules with existential quantifiers, we reformulate composition rules: $(e_x, r_1, e_y) \wedge (e_y, r_1, e_z) \rightarrow (e_x, r_3, e_z)$ is transformed to $(e_x, r_3, e_z) \rightarrow \exists e_y (e_x, r_1, e_y) \wedge (e_y, r_2, e_z)$. For rules involving negation, we calculate

$$\frac{|\{z \mid (e_x, r_1, z) = \text{True} \cap (e_x, r_2, z) = \text{False}\}|}{|\{z \mid (e_x, r, z) \text{ is true}\}|}$$

to identify rules of the form $(e_x, r_1, e_z) \rightarrow \neg(e_x, r_2, e_z)$. In total, we collect 2256 rule candidates at this step.

B.2 Human Verification for Logical Rules

We hired crowdsourcing workers from Amazon Mtruk to verify the plausibility of the rule candidates. We picked up annotators through a qualification test, where worker were asked to annotate 20 carefully-picked rules that are labeled by the author. Only the workers with masters qualification type from English-speaking countries could take part in our qualification test and formal task. The annotators should read the instructions (the full instruction is shown in the box below) thoroughly before beginning working on the task. In the end, seven top-performing workers with at least 80% accuracy were invited to join the formal task. For each natural language expressions of logical rule, we asked the annotator to select one option of its plausibility and justify their option with a short explanation. In practice, we found the explanation sufficient for us to push the workers to provide high-quality annotation and keep track of their performance. Figure 7 shows the screenshot of the annotation interface. We paid 0.30\$ USD for each HIT.

Each rule was labeled by at least two annotators. We filtered out rules to ensure annotators' **perfect agreement** on implausibility, excluding any rule labeled *Contradictory* by any worker. In total, there were approximately 2000 rules verified by the workers.

Our annotation process followed the protocol approved by our university's ethics board, and the author completed an ethics training course before starting the task. Before releasing the task, the author went through all the rule candidates to make sure there were no harmful content. All annotation workers were informed that their contributions would be used for research purposes.

Instructions

This project will present statements about different subjects, such as people, countries, and arts, along with their relationships. You will help us identify how plausible the statement is in a real-world scenario. These statements consist of conditions and outcomes written in the format of “If **conditions**, then **outcomes**,” connecting different facts and events.

For instance, you might see a statement like the following one:

If [person B] is the father of [person A],
then [person A] is the child of [person B].

Your job is to evaluate the plausibility of each statement. In other words, if we assume that the conditions are true, how plausible is it that we can conclude that the outcome is true too?

Concretely, you need to select one of the following options:

- Unrelated: The given conditions are not related to the likelihood of the outcome.
- Contradictory: The conditions and the outcome contradict each other.
- Less likely: The outcome is **less** likely to be true, compared to the case without the conditions.
- More likely: The outcome is **more** likely to be true, compared to the case without the conditions.
- Necessarily true: The outcome is necessarily true, given the conditions are true.

For the above example, we expect that “Necessarily true” is selected. After selecting the option, please write a short explanation (with no fewer than six words) to justify your answer.

Examples

Example of contradictory

If a [team] achieves victory in an [race],
then the [race] cannot have the [team] as a participating entity.

- Reason: A team must participate in a race to achieve victory in that same race.

Example of unrelated

If a [person] has an [profession] as their occupation,
then there exists a [military rank] that they hold, and
this [military rank] is associated with that same [profession].

- Reason: Knowing someone’s occupation does not necessarily mean you know if they have a military rank.

Example of more likely

If a [person] performs a [music work] and
the [music work] is associated with a certain [language],
then the [person] speaks, writes or signs that [language].

- Reason: Singers usually sing in languages they know well, but sometimes they try out languages they don’t use daily.

Example of necessarily true

If a [city] is the capital of an [country] and
that [country] is located in a certain [time zone],
then the [city] is also located in that [time zone].

- Reason: A city shares the same timezone as its country.

Example of less likely

If [person A] is a doctoral student of [person B],
then [person B] is a doctoral student of [person A].

- Reason: According to our common-sense knowledge, it is very unlikely that a PhD advisor pursues another PhD degree with their previous PhD students.

B.3 Licensing

Our dataset will be released under CC-BY-4.0 license. In addition, we will release all the artifacts saved during the data creation process, including the generations from GPT-4o model, plausibility labels and explanation from human annotators.

Figure 7: Mturk annotation task screenshots.

C Data Example in iKnow

Table 4 presents an example from iKnow.

Logical rules and assigned weights	<p>Rule 1: <i>If [TV show] is part of a [TV show series] and the [TV show series] was originally broadcasted on a [broadcaster], then [TV show] was also originally broadcasted on that [broadcaster]. Weight: 1.</i></p> <p>Rule 2: <i>If a [broadcaster] is the original broadcaster of a [TV show] and the [broadcaster] uses a certain [language], then that [language] is the original language of the [TV show]. Weight: 1.</i></p> <p>Rule 3: <i>If a [person] writes the script for [TV show] and they speak and write in [language], then [language] is the original language of the [TV show]. Weight: 9999.</i></p>
Known Facts	<ol style="list-style-type: none"> 1. <i>Game of Thrones, Season 1 is one of the installments in the Overwatch.</i> 2. <i>AMC was the original network for Overwatch.</i> 3. <i>AMC uses Dutch as a primary means of communication.</i> 4. <i>Game of Thrones, Season 1 was written by the screenwriter Peter Jackson.</i> 5. <i>Peter Jackson writes in English.</i> 6. <i>Ridley Scott serves as the screenwriter for The Lost World.</i> 7. <i>Ridley Scott speaks Spanish.</i>
Query Expected Answer	<p>What is the original language of Game of Thrones, Season 1?</p> <p>English</p>
New Facts Expected Rectification	<p>{The Lost World was originally released in Dutch.}</p> <p>{The Lost World was originally released in Spanish.}</p>

Table 4: An example data instance from iKnow includes a knowledge base and two tasks, QA and knowledge rectification, curated from it.

D Experiment Details

We explained the task goals and the functioning of the Markov logic network engine in the system prompts. Table 5 shows the system prompts for both tasks. Next, we added a single in-context example. For CoT prompting, we simply appended "Let's think step by step" to the prompt. We experimented with various system prompt adjustments but observed no significant differences.

We conducted all experiments with open-sourced LLMs using a single Nvidia A100 GPU, with each experiment completing within an average of two days. We used OpenAI's official api to query o1-mini. For each LLM, we tested multiple dataset variants for each task and reported the average performance.

Explanation of MLN	<p>The reasoning process of a Markov Logic Network (MLN) is:</p> <ol style="list-style-type: none"> 1. Represent Knowledge: Use weighted first-order logic rules, where higher weights indicate stronger rules. 2. Ground Formulas: For a given set of predicates, create specific instances (or worlds) to evaluate the truth of these formulas. 3. Calculate Scores: For each possible world (x), compute the score as follows: $\text{score}(x) = \sum_i \exp(w_i \cdot n_i(x))$ where (w_i) are the weights and $n_i(x)$ is the number of true groundings of formula (i) in world (x). <p>By following these steps, you will emulate the MLN’s ability to reason about uncertain knowledge</p>
QA	<p>Read how Markov logic inference works: {{explanation}} Imagine you are a Markov Logic Network engine. Your task is to perform probabilistic reasoning based on a combination of rules and known facts. Provide the most probable answer for the question. Please ensure that you are based solely on the information provided in the prompt, rather than relying on your existing internal knowledge. Wrap the final answer in <answer> and </answer> tags</p>
Knowledge rectification	<p>Read how Markov logic inference works: {{explanation}} Imagine you are a Markov Logic Network engine. You will receive a set of logical rules with associated weights, known facts and a list of new facts. Some new facts may be inconsistent with the rules. Your task is to update up to TWO new facts to make them as consistent with the rules as possible, while minimizing changes and respecting the weights of the rules. Wrap the updated new facts in <update> and </update> tags and separate different facts using <sep>.</p>

Table 5: System prompts for both tasks in experiments with iKnow.

E Additional Experiments on Whether LLMs really “understand” the weights of rules

A potential concern could be that whether LLMs genuinely comprehend the meaning of rule “weights.” To address this, we conducted a targeted experiment testing if LLMs recognize that higher weights correspond to higher priorities. Specifically, we provided the same lists of rules and their corresponding in the same format, then asked LLMs to select the most important rules. Specifically, we provided identical rule lists with their weights presented in the same format and asked the models to select the most important rules. We evaluated the three lowest-performing LLMs from our main experiments, namely Llama3, Qwen2.5, and Phi3, using two weighting schemes (polarized and random) that contain various weights across different rules. Their F1 scores are shown in Table 11. As shown, LLMs can retrieve the most important rules according to their weights with high accuracy. But they fail to incorporate these priorities into reasoning, leading to poor performance on iKnow.

F Additional Experiment Results on Knowledge Rectification

Figure 8 presents the effect of rule weighting schema on the LLM’s performance on knowledge rectification. Similar to what we have observed on the QA task, the performance across different schema is almost the same. Figure 9 reveals that CoT reasoning cannot improve open-sourced LLM’s performance on knowledge rectification task at all.

Error type: Omission of relevant rules	
Input – rules	<p>Rule 1: If a [TV show] was originally broadcasted by a [broadcaster] and the [TV show] originates from a [country], then the [broadcaster] is located in that [country]. Weight: 10.0</p> <p>Rule 2: If a [broadcaster] is associated with a certain [language] and that [language] is widely used in a particular [country], then the [broadcaster] is likely to be in that [country]. Weight: 10.0</p> <p>Rule 3: If a [broadcaster] is licensed to broadcast to a [city] and this [city] is located in a [country], then the [broadcaster] is also licensed to broadcast in that [country]. Weight: 10.0</p> <p>Rule 4: If a [broadcaster] is licensed to broadcast to a particular [city] and that [city] is the capital of a [country], then the [broadcaster] is situated in that [country]. Weight: 10.0</p> <p>...</p>
Input – known facts	<p>The language utilized at Poland is Japanese.</p> <p>Foxtel is officially authorized to serve the broadcast needs of Vancouver.</p> <p>The capital of the Poland is Vancouver.</p> <p>Bretagne contains the administrative territorial entity of Vancouver.</p> <p>The initial transmission of EastEnders was on Foxtel.</p> <p>Poland is the nation that encompasses Sydney.</p> <p>Italy uses Dutch as a primary means of communication.</p> <p>Sydney is part of the administrative jurisdiction of Saskatchewan.</p> <p>The initial transmission of Orange Is the New Black was on Channel One Russia.</p> <p>The country of origin for EastEnders is Poland.</p> <p>Vancouver is situated within the borders of Italy.</p> <p>Foxtel is known by its name in Japanese.</p> <p>Orange Is the New Black was first established in Italy.</p> <p>Vancouver is situated in the administrative entity known as Bretagne.</p> <p>...</p>
Input – queries	In which country can we find Foxtel?
LLM output	<p>Step 1: **Identify the relationship between Foxtel and Vancouver.** - **Fact:** Foxtel is officially authorized to serve the broadcast needs of Vancouver.</p> <p>Step 2: **Determine the status of Vancouver.** - **Fact:** The capital of Poland is Vancouver.</p> <p>Step 3: **Apply Rule 4.** - **Rule 4:** If a [broadcaster] is licensed to broadcast to a particular [city] and that [city] is the capital of a [country], then the [broadcaster] is situated in that [country]. - **Application:** - [broadcaster] = Foxtel - [city] = Vancouver - [country] = Poland</p> <p>Step 4: **Conclude the location of Foxtel.** - Based on Rule 4, since Foxtel is authorized to broadcast in Vancouver, and Vancouver is the capital of Poland, Foxtel is situated in Poland.</p> <p>Final answer: Poland</p>
Groun-truth answer	Italy

Table 6: An example demonstrating how o1-mini fails to incorporate certain relevant rules, highlighted in red, into its reasoning.

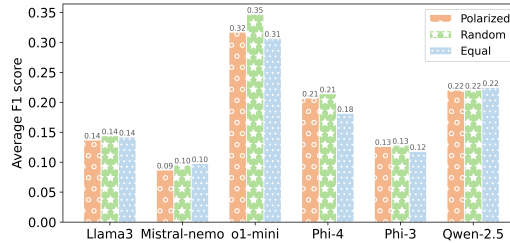


Figure 8: The effect of rule weighting schema on the LLM’s performance on knowledge rectification.

Error type: Hallucination	
Input – rules	<p>If a [person] edits an [film] and the [film] is distributed by an [company], then the [company] is considered to be the employer of the [person]. Weight: 10.0</p> <p>If a [person] is a cast member of a [film/TV show] and the same [film/TV show] is distributed by an [company], then the [company] is considered the employer of the [person]. Weight: 10.0</p> <p>If a [person] is a screenwriter of an [film/TV show] and the [film/TV show] is distributed by an [company], then the [company] is considered the employer of the [person]. Weight: 10.0</p>
Input – known facts	<p>Elijah Wood is the screenwriter of The Vampire Diaries.</p> <p>Amblin Entertainment serves as the distributor for Titanic.</p> <p>Warner Bros. acts as the publisher for the Schindler’s List.</p> <p>Titanic’s raw footage is skillfully shaped into final sequences by Elijah Wood, the film editor.</p> <p>Star Wars reaches audiences through the distribution efforts of Warner Bros.</p> <p>The Vampire Diaries has Amblin Entertainment as its publisher.</p>
Input – queries	For which organization does Elijah Wood work?
LLM output	<p>To solve this, we’ll apply the given rules to the known facts step by step. We’re looking for the organization that is most likely to be Elijah Wood’s employer.</p> <p>1. **Rule 2***: Elijah Wood edits Titanic (fact), and Titanic is distributed by Warner Bros. (fact). Therefore, Warner Bros. is considered Elijah Wood’s employer.</p> <p>...</p>

Table 7: An example showing that Mistral-nemo fails to recall the provided factual statement correctly.

Input – rules	<p>If a [person] is the head of state of a [country] and this [person] speaks, writes or signs in a [language], then this [language] is also widely used in that [country].</p> <p>...</p>
Input – known facts	<p>Betsy DeVos is fluent in Italian.</p> <p>The highest formal authority in Ireland is Betsy DeVos.</p> <p>...</p>
Input – queries	Which language is the common medium for verbal and written communication at Ireland?
LLM output	<p>...</p> <p>2. Now, we need to find out which language is the common medium for verbal and written communication at Ireland. We know that Betsy DeVos is the highest formal authority in Ireland (from the fact "The highest formal authority in Ireland is Betsy DeVos.") and she is fluent in Italian (from the fact "Betsy DeVos is fluent in Italian."). However, there’s no direct evidence that Italian is widely used in Ireland.</p> <p>...</p> <p>Final answer: Cannot be determined with the given information</p>
Ground-truth answer	Italian

Table 8: An example illustrating Mistral-nemo refuse to acknowledge the derived implications that are contradicted to the prior factual knowledge.

Input – rules	<p>...</p> <p>If an [artwork] is part of a [collection] and that [collection] was created by a [person], then that [artwork] was also created by that [person]. Weight: 10.0</p> <p>If [artwork A] is described by [artwork B] and [artwork B] is created by [person], then [artwork A] is also created by that [person]. Weight: 10.0</p> <p>...</p>
Input – known facts	<p>Samuel Taylor Coleridge is the original creator of The National Gallery of Art.</p> <p>The name of The Getty Center is derived from Samuel Taylor Coleridge.</p> <p>Mona Lisa is characterized by Poetics.</p> <p>The National Gallery of Art is named after Johannes Kepler.</p> <p>Secretary of Defense is certified by Canadian Government.</p> <p>The Getty Center was created by Johannes Kepler.</p> <p>Mona Lisa provides a description of Poetics.</p> <p>Poetics belongs to the The Getty Center archival collection.</p> <p>Mona Lisa is housed within the The National Gallery of Art art collection.</p> <p>...</p>
Input – queries	Which individual is acknowledged as the Poetics’s maker?
LLM output	<p>Let’s reason through this step by step:</p> <p>1. We know that “Poetics belongs to the The Getty Center archival collection” and “The Getty Center was created by Johannes Kepler”. This activates rule 4: “If an [artwork] is part of a [collection] and that [collection] was created by a [person], then that [artwork] was also created by that [person].” Therefore, Johannes Kepler is proposed as the creator of Poetics.</p> <p>2. However, we also know that “Mona Lisa is characterized by Poetics” and “Mona Lisa is created by Leonardo da Vinci”. This activates rule 5: “If [artwork A] is described by [artwork B] and [artwork B] is created by [person], then [artwork A] is also created by that [person].” Therefore, Leonardo da Vinci is also proposed as the creator of Poetics.</p> <p>...</p> <p>Final answer: Johannes Kepler</p>
Ground-truth answer	Samuel Taylor Coleridge

Table 9: An example illustrating Mistral-nemo using their learned commonsense knowledge as premise for derivation.

Input – rules	<p>If a [person] is a cast member in an [film/TV show] and the same [film/TV show] is produced by a [company], then the [company] is considered the employer of the [person]. Weight: 10.0</p> <p>...</p>
Input – known facts	<p>The language utilized at University of Zurich is Hindi. United States is where Birth of Venus was created. The narrative of Skyfall unfolds in Chicago. Gladiator was filmed at Chicago. Inception was brought to audiences by Fox 2000 Pictures. Intel Corporation is credited as the maker of The Harry Potter Series. CN Tower is present at the London. Sydney Opera House resides at Paris. Ludwig Van Beethoven owns CN Tower. Sweden is the nation that encompasses Ludwig Van Beethoven. True Detective stars Orson Welles as part of the cast. Ludwig Van Beethoven is credited as the maker of Birth of Venus. Within Kyoto University, German is the commonly used language. Sweden is where The Harry Potter Series was created. Orson Welles is fluent in German. Orson Welles received their education from University of Zurich. Leonardo DiCaprio was educated at Kyoto University. True Detective was brought to audiences by HBO. Inception includes Leonardo DiCaprio in its cast. Leonardo DiCaprio signs in Hindi. ...</p>
Input – queries	For which organization does Leonardo DiCaprio work?
LLM output	<p>To solve this, we'll use the given rules and facts to infer the most probable organization that Leonardo DiCaprio works for. Let's break it down step by step:</p> <ol style="list-style-type: none"> 1. We know that Leonardo DiCaprio was educated at Kyoto University (fact). 2. Kyoto University uses German as its common language (fact). 3. Therefore, Leonardo DiCaprio speaks German (rule 3, weight 10.0). 4. Orson Welles also speaks German (fact). 5. Orson Welles received their education from University of Zurich (fact). 6. University of Zurich uses Hindi as its language (fact). 7. Therefore, Orson Welles speaks Hindi (rule 3, weight 10.0). 8. Leonardo DiCaprio signs in Hindi (fact). 9. This implies that Leonardo DiCaprio and Orson Welles have a common language, which is Hindi. 10. True Detective stars Orson Welles (fact). 11. True Detective was brought to audiences by HBO (fact). 12. Therefore, HBO is a possible employer of Orson Welles (rule 1, weight 10.0). 13. Since Leonardo DiCaprio and Orson Welles share a common language (Hindi), it's likely that they work for the same organization. 14. Thus, the most probable organization that Leonardo DiCaprio works for is HBO. <p>Final answer: HBO</p>
Ground-truth answer	Fox 2000 Pictures

Table 10: An example showing that Mistral-nemo insists on its prior logical knowledge without following the provided one.

LLMs	Polarized	Random
Llama3	98.27	79.17
Qwen2.5	100.00	83.33
Phi3	97.90	61.00

Table 11: LLM performance on rule selection based on their weights

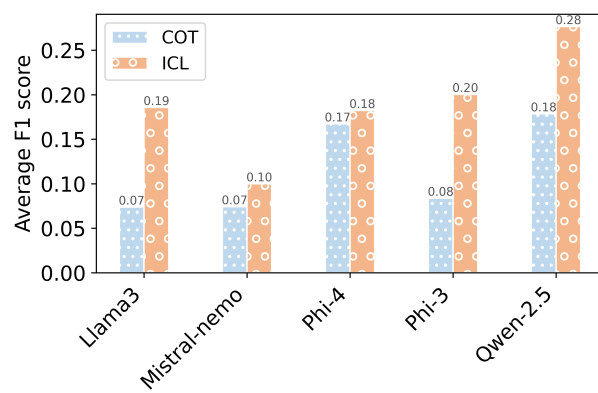


Figure 9: The effect of decoding methods on knowledge rectification.