
Proceedings Paper:
Shafiei, M., Saffari, H. and Moosavi, N.S. (2025) MultiHoax: A dataset of multi-hop false-premise questions. In: Findings of the Association for Computational Linguistics: ACL 2025. Findings of the Association for Computational Linguistics: ACL 2025, 27 Jul - 01 Aug 2025, Vienna, Austria. Association for Computational Linguistics, pp. 10169-10187. ISSN: 0736-587X.

https://doi.org/10.18653/v1/2025.findings-acl.530

MultiHoax: A Dataset of Multi-hop False-Premise Questions

Mohammadamin Shafiei[*1], **Hamidreza Saffari**[*2], **Nafise Sadat Moosavi**[3]
[1]University of Milan, [2]Politecnico di Milano, [3]University of Sheffield
m.shafieiapoorvari@studenti.unimi.it
hamidreza.saffari@mail.polimi.it
n.s.moosavi@sheffield.ac.uk

Abstract

As Large Language Models are increasingly deployed in high-stakes domains, their ability to detect false assumptions and reason critically is crucial for ensuring reliable outputs. False-premise questions (FPQs) serve as an important evaluation method by exposing cases where flawed assumptions lead to incorrect responses. While existing benchmarks focus on single-hop FPQs, real-world reasoning often requires multi-hop inference, where models must verify consistency across multiple reasoning steps rather than relying on surface-level cues. To address this gap, we introduce MultiHoax, a benchmark for evaluating LLMs' ability to handle false premises in complex, multi-step reasoning tasks. Our dataset spans seven countries and ten diverse knowledge categories, using Wikipedia as the primary knowledge source to enable factual reasoning across regions. Experiments reveal that state-of-the-art LLMs struggle to detect false premises across different countries, knowledge categories, and multi-hop reasoning types, highlighting the need for improved false premise detection and more robust multi-hop reasoning capabilities in LLMs.[1]

1 Introduction

In recent years, the evolution of large language models (LLMs) has demonstrated their immense potential across a wide range of natural language processing tasks. However, despite their impressive successes in many domains, their ability to recognize and properly handle false premises in reasoning tasks remains a significant challenge (Hu et al., 2023; Yu et al., 2023). When engaging with false premises, an LLM should be able to identify and reject flawed assumptions rather than proceed as if they were valid.

Which Iranian wrestler won gold in the Men's freestyle 125 kg at the first Olympics when Zahra Nemati was the flag bearer?

Komeil Ghasemi	Ghasem Rezaei	Hassan Yazdani	I do not know

1. The first hop:

At which Olympics did Zahra Nemati carry the flag for the first time?

The answer: Zahra Nemati carried the flag for the first time at the **2016 Rio Olympics**.

2. The second hop:

Which Iranian wrestler won gold in the Men's freestyle 125 kg at the 2016 Rio Olympics?

Falsehood: *No Iranian wrestler won gold in the Men's freestyle 125 kg at the 2016 Rio Olympics.* The only Iranian wrestler to win gold in the Men's freestyle at the 2016 Rio Olympics was Hassan Yazdani in **74 kg**.

Figure 1: A sample MHFPQ from the Sports category related to Iran.

False premises can appear in various forms, such as misleading statements, logically inconsistent claims, or factually incorrect contextual narratives (Leite et al., 2023; Zhuang et al., 2023; Ghosh et al., 2024; Galitsky et al., 2024; Chen and Shu, 2023; Yamin et al., 2024; Zhang et al., 2024b; Csönge, 2015). A common approach to evaluating an LLM's ability to handle false premises is through question-answering (QA) tasks, where a question implicitly contains a misleading or incorrect assumption (Yu et al., 2023; Hu et al., 2023; Kim et al., 2023). For example, a well-performing LLM should recognize that *"Which year did Albert Einstein win the Nobel Prize in Chemistry?"* contains a false premise—Einstein never won a Nobel Prize in Chemistry—and responds appropriately, rather than attempting to provide an incorrect or misleading answer.

Existing benchmarks focus on single-hop false premise detection, where the incorrect assump-

[*] Equal contribution.
[1]The MultiHoax dataset is publicly available at https://github.com/Mamin78/MHFPQ

Findings of the Association for Computational Linguistics: ACL 2025, pages 10169–10187

tion can be identified within a single reasoning step (Yu et al., 2023; Hu et al., 2023; Kim et al., 2023). However, multi-hop reasoning presents a greater challenge, as it requires models to connect multiple pieces of information across multiple inference steps before arriving at an answer (Mavi et al., 2022). Unlike single-hop questions, multi-hop questions (MHQs) require models to derive intermediate conclusions, making them a critical area of research in the evaluation and improvement of LLM reasoning abilities (Mavi et al., 2022; Chen et al., 2024; Yang et al., 2024b; Tang and Yang, 2024; Chakraborty, 2024). The complexity of MHQs arises from the necessity of bridging multiple facts, understanding implicit dependencies between reasoning steps, and maintaining logical consistency throughout the inference process (Mavi et al., 2022).

While MHQs are already challenging, the problem becomes even more difficult when false premises are embedded within the reasoning chain, requiring models not only to answer questions correctly but also to detect and reject incorrect assumptions at intermediate steps. To evaluate LLMs' ability to handle this challenge, we introduce a novel benchmark of Multi-Hop False-Premise Questions (MHFPQs), called **MultiHoax**, that combines the difficulty of false premise questions and multi-hop question answering. MHFPQs are designed to test whether LLMs can detect false premises that appear at one or more reasoning steps rather than simply providing an answer based on a flawed assumption.

Figure 1 shows an example of an MHFPQ from the sports category related to Iran. Answering this question requires a structured reasoning process. First, the model must determine the edition of the Olympics where Zahra Nemati served as Iran's flag bearer for the first time, namely, the 2016 Rio Olympics. Second, it must identify the Iranian wrestler who won the gold medal in the Men's freestyle 125 kg category at those Olympics. This introduces a false premise: no Iranian wrestler won gold in the 125 kg category at the 2016 Olympics. The only Iranian gold medalist in Men's freestyle wrestling at those Games was Hassan Yazdani, who competed in the 74 kg weight class.

To ensure realism and difficulty, we carefully designed the set of possible answer choices in the MHFPQ dataset. Each question includes believable distractors, as possible, that align with historical facts, making the task particularly challenging for LLMs. For instance, in the example from Figure 1, the distractors are real Iranian wrestlers. Moreover, Ghassem Rezaei and Komeil Ghassemi both won gold in freestyle wrestling at the 2012 London Olympics. Hassan Yazdani also won the gold at the 2016 Olympics, but was not competing in the 125 kg division. This ensures that incorrect answers remain plausible while still being contextually invalid. Without detecting the false premise, LLMs may be misled into selecting one of the plausible but incorrect answers, emphasizing the necessity for models to engage in deep contextual reasoning rather than relying on surface-level fact retrieval.

The dataset spans 7 countries and 10 distinct categories, ensuring broad diversity across historical, cultural, and geopolitical contexts. To generate the MHFPQs, we extracted relevant information from Wikipedia pages with the assistance of the Claude model [2]. Each question is paired with three closely related distractors and a single "I do not know" option, which serves as the only valid response in the presence of a false premise. By structuring the benchmark in this way, we aim to assess LLMs' ability to identify and reject incorrect assumptions embedded within multi-hop reasoning chains, rather than merely selecting the most superficially plausible response.

We conduct a comprehensive evaluation of six open-source and closed-source models using our `MultiHoax` dataset of 700 carefully reviewed questions. Our results reveal suboptimal performance across all models, categories, and countries, indicating that current systems struggle with detecting falsehoods in multi-hop reasoning tasks. This resource enables rigorous analysis of multi-hop reasoning under false premises and opens new research directions in question answering that require rejection of complex, embedded assumptions.

2 Related Work

False Information and False Premises. False information is a broad and multifaceted issue, encompassing explicitly misleading claims from social networks (Ma et al., 2022; Rode-Hasinger et al., 2022), blogs (Okazaki et al., 2013), news sources (Long et al., 2017; Qiao et al., 2020; Huang et al., 2023), and online forums (Yu et al., 2023). However, false premises are a distinct challenge since they are not always deliberate misinformation but rather incorrect assumptions that lead to flawed logical conclusions. These premises span various

[2] https://www.anthropic.com/

domains, including historical events (Gabba, 1981; Smith, 2001), religion (Kutlu, 2022), and sports (Dimov, 2021), making them particularly complex to detect.

False-Premise Questions. FPQs assess a model's ability to reason over implicit false assumptions about properties, actions, scope, existence, events, logic, or causality (Hu et al., 2023; Yu et al., 2023). For example, "How many eyes does the sun have?" wrongly assumes the sun has eyes (Hu et al., 2023). Since FPQs require models to identify and reject flawed premises rather than simply answering the question, LLMs often fail to recognize implicit falsehoods, leading to misleading or nonsensical responses (Yu et al., 2023). This inability to reject false premises poses risks for misinformation propagation and model alignment with human expectations.

Early studies on FPQs have focused on obvious falsehoods that humans can easily detect. For instance, Hu et al. (2023) introduced a dataset of simple FPQs, such as "What is the most common color of human's wings?". While earlier LLMs struggled with these, recent versions of the LLMs handle them easily, raising the need to evaluate models on more subtle false premises. Similarly, Kim et al. (2023) proposed a dataset evaluating models on questions containing false or unverifiable assumptions. However, their dataset includes unverifiable claims—statements that may become true over time. For example, "When is Steven Universe Season 5 coming to Hulu?" assumes the event has not yet occurred in order to be false, but this assumption could later become valid. This distinction makes their dataset less suitable for assessing persistent false premises. Another line of research has focused on false-presupposition questions extracted from online forums. Yu et al. (2023) introduced a dataset of FPQs sourced from Reddit, such as "How exactly is current stored in power plants?"—a misleading assumption since electric current is not stored. However, their dataset is primarily limited to scientific and technical topics, lacking the broad factual diversity required for a general evaluation of false premises.

Beyond FPQs, unanswerable questions have also been explored in related research. Zhao et al. (2024) focus on document-grounded unanswerable questions, where a question lacks supporting information within a given document. Their contribution is in evaluating models' ability to reformulate unanswerable questions into answerable ones, rather than directly assessing how LLMs handle false premises in an open-domain setting. Additionally, Lin et al. (2022) study truthful question answering, focusing on questions that elicit misconceptions from humans. Relatedly, Zhang et al. (2024a) and Peng et al. (2024) examine how LLMs handle questions beyond their knowledge, where the correct response should be "I do not know". These studies, however, focus on uncertain or conflicting information rather than inherently false premises.

Multi-Hop False-Premise Reasoning. Most FPQ benchmarks focus on single-hop false premises, yet real-world misinformation often involves multi-step reasoning, making it significantly harder to detect. Multi-hop reasoning questions challenge LLMs by requiring them to connect multiple facts across inference steps. While extensive research has explored general MHQ understanding in LLMs (Yang et al., 2018; Rajabzadeh et al., 2023; Park et al., 2023; Chen et al., 2024), MHQs embedding false premises remain largely unstudied. Existing benchmarks assess multi-hop factual reasoning but do not evaluate whether LLMs can detect and reject false premises within reasoning chains. This gap is critical, as misinformation is rarely an isolated error—false premises are often interwoven across reasoning steps, making them subtle, plausible, and difficult to refute. Evaluating how LLMs handle multi-hop false premises is essential for enhancing their robustness against misinformation and ensuring logical consistency in complex reasoning tasks.

Daswani et al. (2024) attempt to address this with adversarial multi-hop false premise questions, modifying HotpotQA (Yang et al., 2018). Their approach replaces the title of a supporting document with a similar but unrelated distractor, selected based on shared Wikipedia categories. For example, Roger O. Egeberg and Steven K. Galson both fall under American public health doctors, making them interchangeable under this method. However, this technique relies on structured entity swaps rather than embedding implicit falsehoods, making it unclear whether a question contains a truly false premise or is merely unverifiable. Additionally, this approach can lead to unnatural question phrasing, limiting its applicability in assessing real-world false premise detection.

Overall, LLMs struggle with imperfect informa-

tion, conflicting evidence, and questions beyond their knowledge scope (Lin et al., 2022; Zhang et al., 2024a; Peng et al., 2024; Kazemi et al., 2024; Payandeh et al., 2024; Shaier et al., 2024; Park and Lee, 2024; Longpre et al., 2021; Chen et al., 2022). They are also susceptible to distraction by irrelevant details, including false premises, often leading to incorrect, misleading, or fabricated responses (Park and Lee, 2024; Yu et al., 2023; Kim et al., 2023; Hu et al., 2023; Asai and Choi, 2021). These findings underscore a critical gap in evaluating multi-step false premise reasoning, as prior benchmarks focus on single-hop FPQs and adversarial perturbations, failing to capture the complexity of false premises embedded within structured factual reasoning. To address this, we introduce `MultiHoax`, a new benchmark of multi-hop FPQs, designed to assess LLMs' ability to detect and reject false premises within reasoning chains. By including implicit false factual claims across diverse country-related topics, our benchmark enables a more nuanced evaluation of LLMs' ability to detect and reject false premises embedded in complex, multi-hop reasoning chains.

Multi-Regional and Multi-Cultural Resources. Our dataset aligns with multi-cultural and multi-regional NLP research, which examines how LLMs process knowledge across different geographic and cultural contexts. While much prior work has focused on subjective tasks like norms and values (Ziems et al., 2023; Cheng et al., 2023; Fung et al., 2024; Shi et al., 2024; Han et al., 2023; Saffari et al., 2025), recent studies show that multicultural NLP also includes objective knowledge, such as region-specific facts, historical events, and socio-political contexts (Koto et al., 2024; Li et al., 2024; Koto et al., 2023; Son et al., 2024; Kim et al., 2024). This highlights the need to evaluate how LLMs handle factual reasoning across diverse regions, particularly in multi-hop tasks involving false premises.

Despite growing interest in multi-regional fact verification, most work focuses on binary misinformation detection rather than multi-hop false premise reasoning (Thaher et al., 2021; Sabzali et al., 2022; Sheng et al., 2022). A global benchmark for multi-hop false premise detection remains missing, even though factual inaccuracies vary significantly across regions. `MultiHoax` addresses this gap by introducing diverse, context-rich questions across seven countries and ten knowledge domains, enabling evaluation of LLMs' reasoning over embedded falsehoods in culturally grounded settings.

3 MultiHoax Dataset

This section describes the proposed dataset, designed to evaluate multi-hop false premise reasoning across diverse countries, categories, and question types. The dataset structure enables a systematic analysis of how LLMs handle false premises within complex reasoning tasks.

3.1 Dataset Framework

Each *MultiHoax*'s instance consists of the following components:

- **Question and Answer Choices:** Each question contains at least a false premise and is paired with four answer choices, three plausible distractors, and one "I do not know". The latter is positioned randomly to prevent models from exploiting positional bias.
- **False Premise Explanation:** A brief description clarifies why the assumption in the question is incorrect.
- **Country and Category:** The dataset spans seven countries (China, France, Germany, Iran, Italy, the United Kingdom, and the United States) and is categorized into ten knowledge domains, including food, sports, geography, education, history, entertainment, religion, science & technology, arts & literature, and holidays & leisure.
- **Types of False Premises and Answer Types:** Each question is annotated with the type of false premise it contains, following the taxonomy introduced by Hu et al. (2023). The five main types are Property, Event, Entity, and Scope. For example, the question *"Which award did Venkatraman Ramakrishnan receive first: the Shaw Prize in Life Science and Medicine or the Lasker Award?"* falls under the event type, since it implies an event that never happened because he never won the mentioned awards.[3] Furthermore, following Yang et al. (2018), we classify answer types such as person, location, event, number, and common nouns. The answer type shows what the question is asking for.[4]

[3] Table 11 in the appendix presents the distribution of false premise types in the dataset, along with definitions of each type.

[4] Table 12 in the appendix provides a detailed breakdown of answer option types across the dataset, including an example for each type.

- **Multi-Hop Reasoning Type:** Following Mavi et al. (2022), we categorize why a question requires multi-step inference into five types: (1) Named Entity Reasoning: The question requires connecting two facts through an intermediate entity that links them logically, (2) Temporal Reasoning: An intermediate step involves identifying a specific time reference to answer the question correctly, (3) Geographical Reasoning: The reasoning process depends on understanding locations, spatial relationships, or geographic entities, (4) Intersection Reasoning: The answer is determined by an entity that satisfies multiple overlapping conditions, and (5) Comparison Reasoning: The question requires comparing attributes, facts, or values across multiple entities to arrive at the correct conclusion.[5]
- **Wikipedia Grounding:** Each question is linked to a relevant Wikipedia page for factual grounding.

3.2 Wikipedia Document Collection

We developed a pipeline to extract Wikipedia pages relevant to each country and category using ChatGPT-4o's search tool. To select the set of pages, each page was evaluated based on three criteria: relevance to the category, association with the specified country, and existence. "Existence" ensures that the provided link leads to an actual Wikipedia page rather than just an important but undocumented topic. If a page was missing, the model was prompted to complete the list by suggesting alternatives. Ultimately, we collected 15 Wikipedia pages per country and category and extracted their content.

3.3 Question Generation

After collecting relevant documents, we developed a structured process for generating MHFPQs. First, we instructed Claude 3.5 Haiku to extract 15 facts per document using a fact-extraction prompt. Then, based on these facts, we prompted the model to generate MHFPQs. All prompts are detailed in Appendix 6. Due to fundamental differences between Bridge Entity-based MHQs (named, geographical, and temporal entities) and other types (intersection and comparison), we used a separate prompt for each category. We requested three questions from the former and two from the latter but did not enforce specific subtypes (e.g., one intersection, one comparison), as not all documents contained relevant supporting facts. Additionally, we avoided first generating MHQs and then falsifying them, as this could limit the variety of false information types (Daswani et al., 2024) and might not ensure incorrectness.

3.4 Curated Selection and Expert Review

After generating the questions for each category and country, an expert reviewer evaluated the generated questions and selected 10 MHFPQs for each combination of category and country. The selection was guided by a structured, multi-step approach to ensure the quality, validity, and plausibility of the questions.

The first step in the selection process was to verify whether a question adhered to the MH structure. Ensuring diversity in MH question types was a key objective. If a question was not initially formatted as an MHQ but could be converted into one, it passed this initial filter. For instance, the question "What was the name of the Achaemenid ruler who appointed Cyrus as governor of the Median Empire?" is inherently a single-hop reasoning question. However, by introducing an additional reasoning step, such as asking about the ruler's son, it could be transformed into an MHQ: "Who was the son of the Achaemenid ruler who appointed Cyrus as governor of the Median Empire?".

Next, the question needed to contain at least one universally false piece of information—meaning the falsehood had to be global rather than merely incorrect within the context of the associated document. To ensure this criterion was met, only questions that demonstrably contained globally false statements were selected. This verification process followed a rigorous three-step approach. First, the reviewer traced each question back to its corresponding fact or set of facts. Second, these facts were cross-checked against the information found in the relevant document. Finally, the reviewer assessed whether the question was indeed incorrect relative to both the established facts and the document. If the model did not generate enough false information, the reviewer modified the question by introducing falsehoods aligned with the dataset's predefined types of misinformation. However, if adding such falsehoods was not feasible, the question was discarded.

The third step involves explaining why the ques-

[5]Table 18 in the appendix presents the distribution of each multi-hop type in the dataset.

Question	Description	MH Type	Answer Type	FP Type
Which Bundesliga team, Bayern Munich or Borussia Dortmund, has the stadium with the largest seating capacity among clubs that have won the FIFA World Cup?	Clubs do not win the FIFA World Cup, only national teams (like Germany, Brazil, etc.) do. Spotify Camp Nou and Estadio Santiago Bernabéu have the largest capacities among clubs that have won the FIFA Club World Cup, not the FIFA World Cup.	Comparison	Group/Org	Event
Who both served as as a volunteer in the Illinois Militia during the Black Hawk War seeing lots of combat during his tour, and also was among the assassinated presidents of the US?	Abraham Lincoln served as a volunteer in the Illinois Militia April 21, 1832 – July 10, 1832, during the Black Hawk War. However, Lincoln never saw combat during his tour. He was assassinated as well.	Intersection	Person	Property
In which country, besides the U.S. and Italy, was the 1972 American epic gangster film directed by Francis Ford Coppola filmed?	The Godfather (1972) was filmed exclusively in locations around New York City and Sicily, with no scenes shot in other countries.	Named Entity	Location	Event

Table 1: Examples of MHFPQs from MultiHoax.

tion is incorrect. To achieve this, the model's explanation is evaluated based on the previous step's results. If it fails to address the false information, the reviewer provides a more detailed clarification.

The final step involved verifying the plausibility of the answer choices. The reviewer ensured all options were contextually relevant by referring to the corresponding section of the document. If the model's initial options were unrelated—either due to a change in the question's focus or the model's poor performance in generating relevant choices—the reviewer replaced them with more appropriate alternatives.

An example of this procedure is illustrated in Figure 1. The initial question, generated by Claude 3.5 Haiku, was: *At which Olympics did Zahra Nemati carry the flag for the first time?* While this was a factual question, it was not an MHQ, but it could be transformed into one, as shown in Figure 1.

Initially, the question contained no falsehoods—it simply inquired about an event that happened in reality. To introduce false information, we modified the question to ask about an event that did not happen, while adding another reasoning hop. Asking about the Iranian freestyle wrestler who won gold in the 125 kg division at the 2016 Olympics contains a falsehood, as no Iranian wrestler did so at the 2016 Olympics. After these steps, the question was transformed into an MHFPQ, such as: *Which Iranian wrestler won gold in the Men's freestyle 125 kg at the first Olympics when Zahra Nemati was the flag bearer?*

To generate plausible answer choices, we included real Iranian wrestlers. Additionally, we selected the most relevant wrestlers. For example, Komeil Ghasemi won a gold medal in the men's freestyle 120 kg event at the 2012 Summer Olympics. Moreover, Ghasem Rezaei was a former Greco-Roman wrestler and an Olympic gold medalist. Additionally, the only Iranian wrestler to win gold in the Men's freestyle at the 2016 Rio Olympics was Hassan Yazdani in 74 kg. This set of wrestlers provides three relevant and distracting options.

3.5 Secondary False Information Verification

To ensure the accuracy of falsified content, a second round of review was conducted. In this phase, a second reviewer independently examined the description field of each question against the corresponding Wikipedia page to verify the presence of false information. The reviewer categorized each question into one of three outcomes: "There is false information", "There is no false information", and "I cannot tell based on the provided information". [6] Questions confirmed to contain false information were directly included in the final dataset. Those labeled with the second option were double-checked and falsehood was added where required, while those detected with the last option were both improved in terms of clarity and added with false information. In the latter two cases, after the necessary modifications, the reviewer provided feedback to ensure that the question contained clear false information. The dataset was finalized upon completion of this verification process.[7] Table 1 shows examples from our final resource from different

[6] Table 19 in the appendix shows the distribution of the labels across categories.

[7] Table 17 provides the annotation guidelines for this step.

[QUESTION]: 1. [OPTION 1] | 2. [OPTION 2] | 3. [OPTION 3] | 4. [OPTION 4] *Please only provide the answer index.*
Why did you choose "I do not know"? 1. You were uncertain about the question and did not have enough knowledge to answer. 2. You thought the question was wrong and contained false information.

Table 2: Evaluation prompts for multi-hop false premise reasoning where the models are supposed to answer with the index of the option. The top prompt assesses multiple-choice QA, where models may reject false premises by selecting "I do not know". The bottom prompt evaluates whether models correctly justify this choice.

types and countries.

4 Evaluation Setup

The `MultiHoax` dataset serves as the basis for our evaluation of multi-hop false premise reasoning through two primary tasks: (1) closed multiple-choice QA and (2) justification-based verification. In the first task, models were presented with a question and four answer choices, including "I do not know", allowing them to reject the question if they identified a false premise. The prompt used for this step is shown in the first part of Table 2.

In the second task (justification), models that selected "I do not know" are prompted to explain their decision, as shown in the second part of Table 2. This step differentiates between cases where the model lacked knowledge and those where it explicitly identified a false premise. Only responses that both select "I do not know" in the first task and correctly justify it as a false premise in the second task can be considered successful detections of false premises. This justification step enhances the reliability of our evaluation, ensuring that refusal to answer stems from false premise detection rather than general uncertainty.

We additionally explore an alternative output format using structured JSON, where models are asked to produce both an answer and a textual explanation. While this format allows for more expressive reasoning, it can make it harder to unambiguously determine whether a model has identified a false premise, as rejection may be implied rather than explicit. By contrast, the multiple-choice setup includes an explicit "I do not know" option, enabling clearer detection of false-premise rejection. Despite the possibility that this option may simplify the task, it provides a more standardized and reliable evaluation framework.

Models We tested 6 different proprietary and open-source LLMs in our experiments. The models include Claude Sonnet 3.5[8], Gemini-2.0-pro-exp (Team et al., 2023), GPT-4o (Hurst et al., 2024), Qwen2.5-7B-Instruct (Yang et al., 2024a), Llama-3.1-8B-Instruct (Meta et al., 2024), and Deepseek-llm-7b-chat (Bi et al., 2024). All experiments used a zero temperature setting to ensure deterministic responses, with all data collected in February 2025.

5 Results

Model	1st Task	2nd Task
Claude Sonnet 3.5	**0.46**	0.23
Gemini-2.0-pro-exp	0.29	**0.26**
GPT-4o-2024-11-20	0.23	0.25
Qwen2.5-7B-Instruct	0.19	0.03
Llama-3.1-8B-Instruct	0.13	0.01
Deepseek-llm-7b-chat	0.05	0.06

Table 3: Model performance on `MultiHoax`, evaluating multi-hop false premise reasoning in two tasks: (1) one-token multiple-choice QA, where models may reject false premises by selecting "I do not know"; and (2) justification, where models must correctly explain that choice to confirm recognition of a false premise.

Table 3 presents model accuracy on the first group of tasks, which are the first two tasks. The first task evaluates whether models correctly reject false premises by selecting "I do not know". While Claude, which was used during the question-generation phase, demonstrates higher accuracy compared to other models, the models generally struggle to recognize falsehoods in the first task, as they are unable to refuse to answer. Furthermore, although all models show poor performance on this task, open-source models generally exhibit lower accuracy than proprietary ones.

The second task analyzes whether models can justify their "I do not know" responses by correctly identifying the false premise. Here, Gemini slightly outperforms the generator model (Claude),

[8] https://www.anthropic.com/

Category / Model	Claude	Gemini	GPT	Qwen	Llama	Deepseek	Avg
Science and Technology	0.458	0.329	0.286	0.215	0.143	0.430	**0.310**
Entertainment	0.429	0.286	0.172	0.129	0.115	0.000	0.189
Education	0.529	0.343	0.200	0.215	0.129	0.072	0.248
Art and Literature	0.458	0.258	0.172	0.072	0.086	0.043	0.182
Food	0.529	0.300	0.186	0.200	0.158	0.072	0.241
Religion	0.543	0.200	0.200	0.158	0.100	0.058	0.210
Sports	0.443	0.415	0.315	0.300	0.186	0.058	0.286
Holiday, Celebrations, and Leisure	0.580	0.286	0.315	0.229	0.100	0.072	0.264
Geography	0.343	0.243	0.215	0.186	0.172	0.058	0.203
History	0.343	0.286	0.258	0.172	0.158	0.058	0.213
Avg	**0.466**	0.295	0.232	0.188	0.135	0.092	–

Table 4: Accuracy of models across knowledge categories in MultiHoax.

Country / Model	Claude	Gemini	GPT	Qwen	Llama	Deepseek	Avg
China	0.49	0.32	0.23	0.19	0.17	0.02	0.24
France	0.45	0.29	0.23	0.21	0.15	0.06	0.23
Germany	0.47	0.31	0.30	0.16	0.10	0.06	0.23
Iran	0.47	0.34	0.25	0.19	0.13	0.05	0.24
Italy	0.52	0.31	0.24	0.17	0.12	0.03	0.23
UK	0.47	0.31	0.22	0.25	0.17	0.08	**0.25**
USA	0.37	0.18	0.15	0.14	0.10	0.07	0.17
Avg	**0.46**	0.29	0.23	0.19	0.13	0.06	–

Table 5: Accuracy of models across countries in MultiHoax.

Multi-hop Type	Claude	Gemini	GPT	Qwen	Llama	DeepSeek	Avg
Comparison	0.593	0.373	0.271	0.153	0.153	0.101	**0.274**
Geographical	0.439	0.367	0.235	0.224	0.173	0.082	0.253
Intersection	0.466	0.271	0.227	0.171	0.112	0.036	0.214
Named	0.445	0.283	0.191	0.197	0.133	0.035	0.214
Temporal	0.437	0.261	0.277	0.193	0.143	0.067	0.230
Avg	**0.476**	0.311	0.240	0.188	0.143	0.064	–

Table 6: Accuracy of models across different multi-hop reasoning types in MultiHoax.

but overall accuracy remains low across all models, indicating persistent difficulty in recognizing false premises.

Table 4 presents model performance across ten knowldege categories. Accuracy is highest in science and technology, while art and literature and entertainment rank lowest, reflecting variation in LLMs' domain-specific factual competence. Notably, proprietary models consistently outperform open-source models, though performance remains low across all categories. Interestingly, Claude's dominance is most pronounced in high-information domains like education and food, while other models exhibit irregular performance trends. These results highlight significant disparities in model strengths across different knowledge areas.

Table 5 analyzes accuracy across countries. The results indicate close accuracy levels, suggesting that models struggle to detect falsehoods regardless of the country. The lower accuracy in some countries can be a result of more challenging questions, which can suggest that the generator model was better able to design challenging questions for those countries.

Table 6 presents model accuracy across multi-hop reasoning types. While comparison-based questions show slightly higher accuracy, the differences across reasoning types are modest, and overall performance remains low. This suggests that all multi-hop reasoning categories pose significant challenges to current models.[9]

Finally, we evaluate models using a structured JSON output format that encourages explicit reasoning. This setting uses the prompt shown in Ta-

[9]Tables 20 and 21 in the appendix provide model accuracy across different false premise and answer types.

[QUESTION]:
Return the answer in a JSON format with a brief 'explanation' and an 'answer' field.
Here is an example output:
{
"explanation": "The University of Tehran Press published the works of Hafez when it was founded."
"answer": "Hafez"
}
ONLY RETURN THE JSON.

Table 7: Evaluation prompts for the JSON output structure format.

Model	Accuracy
Claude	**0.46**
GPT	0.20
Gemini	0.18
Qwen	0.22

Table 8: Accuracy of models when asked to provide structured JSON outputs with both answer and explanation fields.

ble 7. Table 8 presents the results of JSON output for the top-performing models from the previous task. The results indicate accuracy levels comparable to those observed in the multiple-choice setting. This suggests that models consistently struggle to detect embedded falsehoods in the `MultiHoax` dataset, regardless of output format or reasoning constraints.

6 Conclusion

We introduce a novel class of multi-hop false-premise questions (MHFPQs), combining the complexities of multi-hop reasoning and false-premise detection. To support this research, we present `MultiHoax`, the first FPQ resource that spans multiple countries and knowledge domains, enabling evaluation of LLMs' ability to navigate multi-step falsehoods in diverse contexts. Our dataset provides a comprehensive evaluation framework, spanning seven countries and ten knowledge categories, allowing for a detailed analysis of how LLMs handle false premises across diverse topics and regions. Unlike prior FPQ datasets, MHFPQs require deeper reasoning, as falsehoods are not immediately apparent but emerge through multi-step inference. `MultiHoax` establishes a novel evaluation setting at the intersection of multi-hop reasoning and false-premise detection, revealing persistent limitations in current LLMs. By targeting assumption-level failures across diverse domains and regions, it provides a rigorous benchmark for advancing models that can reason robustly under implicit factual errors.

Limitations

Our dataset has some limitations. While we have aimed to include a diverse range of countries with varying levels of resource availability, there are still opportunities to incorporate additional countries from other regions worldwide.

Second, our category set, which consists of 10 categories, could be expanded as scholars explore knowledge across various areas. While we have focused on a set of categories suitable for factual, objective questions, other potential categories could be included. Additionally, subjective questions, such as "Who first imported the most popular type of ingredient to China?" could be considered. This would be an MHQ, as it requires identifying both the most popular ingredient and the first person to import it. However, determining the most popular ingredient is subjective, and the question becomes MHFP if the ingredient was never imported to China. While subjective questions are feasible, reviewing them differs significantly from the process for factual objective questions.

Furthermore, while we include questions associated with a diverse set of countries, we do not have the translation of these questions in the local languages of these countries, except from the United States and the United Kingdom. Future research can be collecting questions in the local language of such countries or translating ours to those languages.

Finally, while our current resource presents a significant challenge to different LLMs, and even the best models struggle with our tasks, the rapid progress of LLMs may make our dataset less difficult over time. As models improve, we may need to update our resources to introduce more challenging tasks that better test LLMs' reasoning abilities.

Ethical Considerations

We aim to provide a set of factually incorrect questions requiring multiple reasoning steps to challenge various models' ability to detect falsehoods.

To compile our dataset, we utilized LLMs to generate potential questions, which greatly facilitated the process. However, this approach may introduce biases, as LLMs are more knowledgeable about certain countries than others. For instance, the USA's lower accuracy could be attributed to LLMs having more in-depth knowledge of U.S. facts, allowing them to craft more challenging questions. Although we used Wikipedia documents to mitigate this bias, it cannot be entirely eliminated. In future iterations, we plan to diversify the question types, incorporating topics focused on values and norms rather than just factual knowledge, and we aim to minimize reliance on LLMs for question generation.

References

Akari Asai and Eunsol Choi. 2021. Challenges in information-seeking QA: Unanswerable questions and paragraph retrieval. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1492–1504, Online. Association for Computational Linguistics.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Abir Chakraborty. 2024. Multi-hop question answering over knowledge graphs using large language models. *arXiv preprint arXiv:2404.19234*.

Canyu Chen and Kai Shu. 2023. Can llm-generated misinformation be detected? In *The Twelfth International Conference on Learning Representations*.

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*.

Myra Cheng, Tiziano Piccardi, and Diyi Yang. 2023. Compost: Characterizing and evaluating caricature in llm simulations. *arXiv preprint arXiv:2310.11501*.

Tamás Csönge. 2015. Moving picture, lying image: Unreliable cinematic narratives. *Acta Universitatis Sapientiae, Film and Media Studies*, (10):89–104.

Ashwin Daswani, Rohan Sawant, and Najoung Kim. 2024. Syn-QA2: Evaluating false assumptions in long-tail questions with synthetic qa datasets. *arXiv preprint arXiv:2403.12145*.

Petko Dimov. 2021. Recognition of fake news in sports. , 29(4s):18–27.

Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. Massively multi-cultural knowledge acquisition & lm benchmarking. *arXiv preprint arXiv:2402.09369*.

Emilio Gabba. 1981. True history and false history in classical antiquity. *The Journal of Roman Studies*, 71:50–62.

Boris Galitsky, Anton Chernyavskiy, and Dmitry Ilvovsky. 2024. Truth-o-meter: Handling multiple inconsistent sources repairing llm hallucinations. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2817–2821.

Bishwamittra Ghosh, Sarah Hasan, Naheed Anjum Arafat, and Arijit Khan. 2024. Logical consistency of large language models in fact-checking. *arXiv preprint arXiv:2412.16100*.

Seungju Han, Junhyeok Kim, Jack Hessel, Liwei Jiang, Jiwan Chung, Yejin Son, Yejin Choi, and Youngjae Yu. 2023. Reading books is great, but not if you are driving! visually grounded reasoning about defeasible commonsense norms. *arXiv preprint arXiv:2310.10418*.

Shengding Hu, Yifan Luo, Huadong Wang, Xingyi Cheng, Zhiyuan Liu, and Maosong Sun. 2023. Won't get fooled again: Answering questions with false premises. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5626–5643, Toronto, Canada. Association for Computational Linguistics.

Kung-Hsiang Huang, Kathleen McKeown, Preslav Nakov, Yejin Choi, and Heng Ji. 2023. Faking fake news for real fake news detection: Propaganda-loaded training data generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14571–14589, Toronto, Canada. Association for Computational Linguistics.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Mehran Kazemi, Quan Yuan, Deepti Bhatia, Najoung Kim, Xin Xu, Vaiva Imbrasaite, and Deepak Ramachandran. 2024. Boardgameqa: A dataset for natural language reasoning with contradictory information. *Advances in Neural Information Processing Systems*, 36.

Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. 2024. CLIcK: A benchmark dataset of cultural and linguistic intelligence in Korean. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3335–3346, Torino, Italia. ELRA and ICCL.

Najoung Kim, Phu Mon Htut, Samuel R. Bowman, and Jackson Petty. 2023. $(QA)^2$: Question answering with questionable assumptions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8466–8487, Toronto, Canada. Association for Computational Linguistics.

Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. 2023. Large language models only pass primary school exams in Indonesia: A comprehensive test on IndoMMLU. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12359–12374, Singapore. Association for Computational Linguistics.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024. ArabicMMLU: Assessing massive multitask language understanding in Arabic. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5622–5640, Bangkok, Thailand. Association for Computational Linguistics.

Mahmut Kutlu. 2022. Analysis of false religious posts on social media. *Ahi Evran Akademi*, 3(2):14–30.

João A Leite, Olesya Razuvayevskaya, Kalina Bontcheva, and Carolina Scarton. 2023. Detecting misinformation with llm-predicted credibility signals and weak supervision. *arXiv preprint arXiv:2309.07601*.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024. CMMLU: Measuring massive multitask language understanding in Chinese. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11260–11285, Bangkok, Thailand. Association for Computational Linguistics.

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

Yunfei Long, Qin Lu, Rong Xiang, Minglei Li, and Chu-Ren Huang. 2017. Fake news detection through multi-perspective speaker profiles. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 252–256, Taipei, Taiwan. Asian Federation of Natural Language Processing.

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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Guanghui Ma, Chunming Hu, Ling Ge, and Hong Zhang. 2022. Open-topic false information detection on social networks with contrastive adversarial learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2911–2923, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Vaibhav Mavi, Anubhav Jangra, and Adam Jatowt. 2022. A survey on multi-hop question answering and generation. *arXiv preprint arXiv:2204.09140*.

AI Meta, 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*, 2.

Naoaki Okazaki, Keita Nabeshima, Kento Watanabe, Junta Mizuno, and Kentaro Inui. 2013. Extracting and aggregating false information from microblogs. In *Proceedings of the Workshop on Language Processing and Crisis Information 2013*, pages 36–43, Nagoya, Japan. Asian Federation of Natural Language Processing.

Jinyoung Park, Ameen Patel, Omar Zia Khan, Hyunwoo J Kim, and Joo-Kyung Kim. 2023. Graph-guided reasoning for multi-hop question answering in large language models. *arXiv preprint arXiv:2311.09762*.

Seong-Il Park and Jay-Yoon Lee. 2024. Toward robust RALMs: Revealing the impact of imperfect retrieval on retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 12:1686–1702.

Amirreza Payandeh, Dan Pluth, Jordan Hosier, Xuesu Xiao, and Vijay K. Gurbani. 2024. How susceptible are LLMs to logical fallacies? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8276–8286, Torino, Italia. ELRA and ICCL.

Zhiyuan Peng, Jinming Nian, Alexandre Evfimievski, and Yi Fang. 2024. Scopeqa: A framework for generating out-of-scope questions for rag. *arXiv preprint arXiv:2410.14567*.

Yu Qiao, Daniel Wiechmann, and Elma Kerz. 2020. A language-based approach to fake news detection through interpretable features and BRNN. In *Proceedings of the 3rd International Workshop on Rumours and Deception in Social Media (RDSM)*, pages 14–31, Barcelona, Spain (Online). Association for Computational Linguistics.

Hossein Rajabzadeh, Suyuchen Wang, Hyock Ju Kwon, and Bang Liu. 2023. Multimodal multi-hop question answering through a conversation between tools and efficiently finetuned large language models. *arXiv preprint arXiv:2309.08922*.

Samyo Rode-Hasinger, Anna Kruspe, and Xiao Xiang Zhu. 2022. True or false? detecting false information on social media using graph neural networks. In *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, pages 222–229, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Maryam Sabzali, Majid Sarfi, Mostafa Zohouri, Tahereh Sarfi, and Morteza Darvishi. 2022. Fake news and freedom of expression: An iranian perspective. *Journal of Cyberspace Studies*, 6(2):205–218.

Hamidreza Saffari, Mohammadamin Shafiei, Donya Rooein, Francesco Pierri, and Debora Nozza. 2025. Can I introduce my boyfriend to my grandmother? evaluating large language models capabilities on Iranian social norm classification. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6060–6074, Albuquerque, New Mexico. Association for Computational Linguistics.

Sagi Shaier, Ari Kobren, and Philip V. Ogren. 2024. Adaptive question answering: Enhancing language model proficiency for addressing knowledge conflicts with source citations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17226–17239, Miami, Florida, USA. Association for Computational Linguistics.

Qiang Sheng, Juan Cao, H Russell Bernard, Kai Shu, Jintao Li, and Huan Liu. 2022. Characterizing multi-domain false news and underlying user effects on chinese weibo. *Information Processing & Management*, 59(4):102959.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Raya Horesh, Rogério Abreu de Paula, Diyi Yang, et al. 2024. Culturebank: An online community-driven knowledge base towards culturally aware language technologies. *arXiv preprint arXiv:2404.15238*.

Andrew F Smith. 2001. False memories: The invention of culinary fakelore and food fallacies. In *Food and the Memory: Proceedings of the Oxford Symposium on Food and Cookery*, pages 254–260.

Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2024. Kmmlu: Measuring massive multitask language understanding in korean. *arXiv preprint arXiv:2402.11548*.

Yixuan Tang and Yi Yang. 2024. Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries. *arXiv preprint arXiv:2401.15391*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Thaer Thaher, Mahmoud Saheb, Hamza Turabieh, and Hamouda Chantar. 2021. Intelligent detection of false information in arabic tweets utilizing hybrid harris hawks based feature selection and machine learning models. *Symmetry*, 13(4):556.

Khurram Yamin, Shantanu Gupta, Gaurav R Ghosal, Zachary C Lipton, and Bryan Wilder. 2024. Failure modes of llms for causal reasoning on narratives. *arXiv preprint arXiv:2410.23884*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. 2024b. Do large language models latently perform multi-hop reasoning? *arXiv preprint arXiv:2402.16837*.

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

Xinyan Yu, Sewon Min, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. CREPE: Open-domain question answering with false presuppositions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10457–10480, Toronto, Canada. Association for Computational Linguistics.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024a. R-tuning: Instructing large language models to say 'I don't know'. 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 7113–7139, Mexico City, Mexico. Association for Computational Linguistics.

Wenyuan Zhang, Jiawei Sheng, Shuaiyi Nie, Zefeng Zhang, Xinghua Zhang, Yongquan He, and Tingwen Liu. 2024b. Revealing the challenge of detecting

character knowledge errors in llm role-playing. *arXiv preprint arXiv:2409.11726*.

Wenting Zhao, Ge Gao, Claire Cardie, and Alexander M Rush. 2024. I could've asked that: Reformulating unanswerable questions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4207–4220, Miami, Florida, USA. Association for Computational Linguistics.

Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. 2023. Toolqa: A dataset for llm question answering with external tools. In *Advances in Neural Information Processing Systems*, volume 36, pages 50117–50143. Curran Associates, Inc.

Caleb Ziems, Jane Dwivedi-Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2023. Normbank: A knowledge bank of situational social norms. *arXiv preprint arXiv:2305.17008*.

Appendix A: Prompts

In this section, we provide the different prompts that we leveraged to collect the Wikipedia pages, and also prompts to generate the questions.

Wikipedia Document Retrieval Prompt

In this section, we first define each category. Then, we used them in order to search for the relevant documents of each category using ChatGPT-4o's search using the prompt in Table 9.

Food

- **Cuisine**: Signature dishes, cooking styles, traditional meals.
- **Ingredients**: Locally grown spices, crops, and special ingredients.
- **Drinks**: Popular beverages, traditional teas, or alcoholic drinks.

Sports

- **National and Popular Sports**: Widely played or watched sports in the country and Official sports of a country.
- **Athletes**: Famous sportspeople or Olympic medalists.
- **Tournaments and Sports Venues**: Major leagues, championships, or cups, as well as iconic stadiums, arenas, or tracks.

Education

- **Education System AND Literacy**: Structure (primary, secondary, higher education) AND Efforts to promote literacy or improve access to education.
- **Schools and Universities AND Curriculum**: Prestigious or historic institutions AND Subjects emphasized or unique courses.
- **Famous Educators**: Scholars, reformers, or pioneers in education.

Holidays/Celebrations/Leisure

- **National Holidays**: Independence days, constitution days, or memorials.
- **Festivals**: Cultural, religious, or seasonal festivals.
- **Others**: Other topics related to Holidays/Celebrations/Leisure.

History

- **Historical Figures**: Leaders, revolutionaries, empires and kingdoms, or intellectuals.
- **Important Events**: Battles, treaties, or turning points in history.
- **Landmarks**: Historical monuments or UNESCO heritage sites.

Geography

- **Natural Features AND Resources**: Mountains, rivers, lakes, and deserts AND Natural resources, agriculture, or energy production.
- **Cities AND Regions**: Capitals, major cities, or urban landmarks AND Administrative divisions or cultural regions.
- **Geopolitics**: Borders, neighbors, or disputed territories.

Science and Technology

- **Scientists**: Modern renowned scientists.
- **Engineering**: Famous modern constructions, bridges, or technology.
- **Others**: Other related topics to Science and Technology like Medical Breakthroughs, Research Centers, Computing Pioneers, Green Technology, Digital Platforms, and Communications.

We are developing a dataset of factual questions and aiming to collect questions across various categories and countries. Currently, we are identifying relevant important Wikipedia pages for each category and country. Please gather Wikipedia pages related to **[CATEGORY]** for the following countries:

- China
- France
- Germany
- Italy
- United States
- United Kingdom
- Iran

Please provide **15** distinct Wikipedia pages related to **[CATEGORY]** for the mentioned countries, based on the following category definition below.
`[DEFINITION OF THE CATEGORY]`
ONLY PROVIDE THE PAGE NAMES LINKED TO THE PAGE WITHOUT ANY EXPLANATION.

Table 9: The prompt for searching for related documents using ChatGPT4o.

Arts and Literature

- **Writers**: Prominent authors, poets, or playwrights.
- **Books**: National epics, famous novels, or historical documents.
- **Artists**: Prominent artists.

Religion

- **Religions and Religious Practices**: Popular religions and Worship styles or Religious rituals.
- **Holy Sites**: Temples, churches, mosques, or pilgrimage locations.
- **Others**: Religious Leaders, Religious Festivals, or Sacred Texts.

Entertainment

- **Cinema and TV**: National cinema, famous directors, popular movies, or actors.
- **Music**: Traditional music styles, Musicians, or iconic bands.
- **Others**: Other topics related to entertainment like theater, gaming, festivals related to entertainment, or media.

These category definitions were used in the prompt in the Table 9 to get 15 related pages for each country.

Extract 15 facts from the document.
Document Content: **DOCUMENT CONTENT**

Table 10: Fact extraction prompt.

Fact Extraction Prompt

To extract facts from each document, we leveraged a simple brief prompt. In the fact extraction prompt, we define the number of facts to be extracted, the name of the document, and the document content. The prompt is provided in the Table 10.

Question Generation Prompt

For question generation, we designed two distinct prompts corresponding to the two major types of MHQs: Entity-based and Intersection & Comparison Combination. Each prompt consists of three parts: an introduction to the task, an explanation of the specific MHQ type, and a set of generation

Type	Percentage (%)
Intersection	36
Named-entity	25
Temporal-entity	17
Geographical-entity	14
Comparison	8

Table 11: Distribution of MH types in the dataset.

Type	Percentage (%)
Person	44
Group/Org	13
Location	8
Number	7
Language or Nationality or Country	6
Date	6
Common noun	5
Food	4
Artwork	3
Event	2
Other proper noun	2

Table 12: Distribution of Answer types in the dataset.

rules. The first section, shown in the Table 13, and the last section, shown in the Table 16 are the same in both generation prompts, but the second part, the definition of the specific MHQ types, is different.

The tables 14 and 15 show the two different second sections.

Appendix B: Annotation Guideline

Annotation Guideline

The annotation guideline, shown in the Table 17 was used to familiarize false information reviewers with the dataset structure and their tasks. The guideline first defines what an FPQ is, then explains how questions are structured in our dataset, and finally outlines the reviewing task.

Annotators details

Following the initial round of annotation by the authors, the dataset was divided into three parts for the second review phase. We then recruited three university student annotators—one female and two male—each paid £12.21 per hour for ten hours of work.

Appendix C: Statistics

In this section, we provide further statistical information regarding the dataset.

Table 11 provides the distribution of the various types of MHQs.The majority of the resource consists of bridge entity-based questions, while the intersection type is largely included.

Table 18 shows the distribution of the various types of false information included in the questions. We include four types in our resources. The description of types is also provided in Table 18.

Moreover, Table 12 provides the details of answer type occurrence over the resource, the person type as the major type.

Regarding the annotation details, Table 19 contains the number of times the second reviewer chose each possible label for the questions.

The table 20 provides the detailed accuracy of the models across different types of answers. As shown, event, location, and number have the lowest accuracies, suggesting that the models find these types more challenging than the rest while they generally show poor performance.

Tables 6 and 21 also show the detailed accuracies for different types of MHQs and FPQs.

We are developing a dataset of counterfactual multi-hop reasoning false-premise questions (MHFPQs). Multi-hop reasoning questions require retrieving and connecting multiple pieces of information across two or more logical steps to derive the final answer. We have seen what a multi-hop question is.
Now, let's focus on creating questions with false premises—where in the question there exists a false assumption that makes the question wrong and impossible to answer correctly. MHFPQs have similar types to Multi-hop but with false premises.
Here are some examples: Note that these are only examples, DO NOT include them in your answers.

Table 13: The first part of the generation prompt.

- **Temporal Multi-Hop Reasoning**
  Example: Who was the president of Iran in the year in which Ehsan Rouzbahani won the Olympics Bronze medal in Tokyo?

  Description: Ehsan Rouzbahani did not compete in the Tokyo Olympics, making it impossible for him to win a (bronze) medal.

  Reasoning Steps:

  1. The year when Ehsan Rouzbahani won the Olympic bronze medal in Tokyo.
  2. The president of Iran at that time.

- **Geographical Multi-Hop Reasoning**
  Example: Which football team with the most championships in the territory Alexander the Great conquered before turning 18?

  Description: Alexander the Great did not conquer any territory before turning 18.

  Reasoning Steps:

  1. The territory Alexander the Great conquered before turning 18.
  2. Football team with the most championships in that territory.

- **Named Entity Multi-Hop Reasoning**
  Example: What is the camera brand used by Spielberg when filming his Academy Award-winning student film at USC?

  Description: Spielberg never attended USC and didn't win an Academy Award as a student.

  Reasoning Steps:

  1. The Spielberg's Academy Award-winning student film at USC.
  2. The camera brand used for that film.

Table 14: The second part of the generation prompt for the first group.

- **Intersection-Type Multi-Hop Reasoning**
  Example: Which architect both designed the golden-domed Old Basilica and incorporated Aztec symbols in its facade in 1695?

  Description: The Old Basilica did not have Aztec symbols. It only had yellow and blue Talavera mosaics.

  Reasoning Steps:

  1. The architect who designed the golden-domed Old Basilica.
  2. The architect who incorporated Aztec symbols in the Old Basilica's facade in 1695.

- **Comparison-Type Multi-Hop Reasoning**
  Example: Which of the Chinese and the Germans first invented sauerkraut in the 18th century?

  Description: Sauerkraut was not invented in the 18th century, and it existed much earlier.

  Reasoning Steps:

  1. Did the Chinese invented sauerkraut in the 18th century?
  2. Did Germans invented sauerkraut in the 18th century?

Note: In Comparison type, multi-hop reasoning questions, none of the entities satisfies the condition.

Table 15: The second part of the generation prompt for the second group.

Your task is to extract false premise multi-hop questions FROM THE FACTS PROVIDED. Here are the instructions:

- The structure of the question should be like the given structures, but the content can be different.
- False premises are implicitly embedded within the questions. Also, false premises must not be obvious.
- Provide 3 relevant, engaging, and realistic options.
- Questions should be based on the provided FACTS from the specific country.
- Focus is exclusively on verifiable factual claims, avoiding cultural norms or subjective topics.
- For each question, a clear explanation must be provided:
- Identifying the false premise.
- Clarifying the actual truth.
- If it is not possible to design questions from all the types, you can only focus on most probable ones.

Return each question in the following format:
<false_premise_multi_hop_question> | <first_option> | <second_option> | <third_option> | <description> | <reasoning_steps> | <type_of_multi_hop>

Table 16: The third part of the generation prompt.

False-premise Questions: We have a set of false-premise questions. A false premise question is a question with at least one piece of false information. A simple example of false-premise questions can be "How many eyes does the sun have?" Such simple questions include false information that is easily detectable by humans. More challenging false-premise questions, which are the target of our experiment, have false pieces of information that are difficult for non-expert humans to detect.

Question: During which time, 1985 to 1990 or 1995 to 2010, did CERN's affiliates win more Nobel prizes in physics?

1. 1985 to 1990
2. 1995 to 2010
3. In both durations, CERN's affiliates won only 1 Nobel prize
4. I don't know

Explanation: In none of the durations, CERN's affiliates won a Nobel prize. 1984, 1992, and 2013 are the years when CERN's affiliates won the award.
As you can see, such false information types are not detectable unless the person knows about the history of the mentioned institute, which is not the case for non-experts.
Format of Data: In the dataset, there are a number of fields, like the following example.
Question: What is the name of the new Humanistic Buddhist organization that was established in Beijing in the 2000s to promote the revival of Vajrayana Buddhism in China? Options:

1. Cíjì
2. Huácáng Zōngmén
3. Zhēnfó Zōng
4. I don't know

Description: The question contains a false premise that a new Humanistic Buddhist organization was established in Beijing in the 2000s to promote the revival of Vajrayana Buddhism. According to the facts, the Humanistic Buddhist movement in China is associated with organizations like Cíjì, which has been working in mainland China since 1991, not a new organization focused on Vajrayana Buddhism.
Wikipedia: https://en.wikipedia.org/wiki/Buddhism_in_China
Your task: You are supposed to read the question and options, and then check the provided description, which is the description of why the question includes false information. Afterward, you are supposed to label these questions after checking if there is any false information in them or not. "*There is false information*", "*There is no false information*", and "*I cannot tell based on the provided information*" are the possible options. You need to visit the Wikipedia page related to each question and check false information based on that **Wikipedia** page.

- If you choose "*There is no false information*", then you are supposed to explain why you have chosen this. For example, in the above case, if you choose "There is no false information", then an example explanation can be "CERN's affiliates won the Nobel prize in 1986 making the question true".
- If you choose "*I cannot tell based on the provided information*", you must also explain the ambiguity or the problem you have in verifying the question in the explain column.
- If you choose "*There is false information*", then there is no need to explain.

Keep the explanation clear, simple, and concise.

Table 17: Reviewing guidelines.

Type	Description	Percentage (%)
Event	The event didn't happen in history.	49
Property	The entity does not have the property.	36
Scope	A fact is not valid in the scope.	13
Entity	The entity cannot exist.	2

Table 18: Distribution of FP types in the dataset.

File	**False Information**	**Cannot Tell**	**No False Information**
Science and technology	49	12	9
Entertainment	66	3	1
Education	65	5	0
Art and Literature	59	11	0
Food	62	6	2
Religion	59	6	5
Sports	56	8	6
Holiday, Celebrations, and Leisure	52	4	14
Geography	51	3	16
History	57	0	13
Sum	576	57	67

Table 19: Analysis of false information inclusion across different categories based on the second phase review.

Answer Type	**Claude**	**Gemini**	**GPT**	**Qwen**	**Llama**	**DeepSeek**	**Avg**
Artwork	0.480	0.360	0.280	0.160	0.080	0.040	0.233
Common Noun	0.459	0.351	0.243	0.243	0.162	0.162	0.270
Date or Time Period	0.600	0.450	0.375	0.275	0.200	0.025	0.321
Event	0.429	0.143	0.214	0.143	0.071	0.000	0.167
Food	0.577	0.308	0.192	0.192	0.038	0.077	0.231
Group or Org	0.494	0.205	0.241	0.181	0.108	0.060	0.215
Language/Nationality/Country	0.595	0.238	0.238	0.167	0.167	0.095	0.250
Location	0.339	0.271	0.186	0.136	0.136	0.034	0.184
Number	0.333	0.333	0.208	0.146	0.167	0.000	0.198
Other Proper Noun	0.417	0.250	0.167	0.167	0.167	0.167	0.223
Person	0.453	0.296	0.219	0.193	0.135	0.045	0.223
Avg	0.477	0.278	0.226	0.185	0.125	0.066	0.226

Table 20: Accuracy of models across different answer types in `MultiHoax`.

FP Type	**Claude**	**Gemini-2.0**	**GPT-4**	**Qwen2.5**	**Llama-3.1**	**DeepSeek-7B**	**Avg**
Entity	0.778	0.510	0.333	0.278	0.278	0.111	0.381
Event	0.451	0.309	0.209	0.197	0.151	0.062	0.230
Property	0.478	0.251	0.247	0.183	0.112	0.048	0.220
Scope	0.400	0.311	0.244	0.144	0.111	0.022	0.205
Avg	**0.527**	0.345	0.258	0.201	0.163	0.061	0.259

Table 21: Accuracy of models across different false premise (FP) types in `MultiHoax`.