

---

# CONFLICTBANK: A Benchmark for Evaluating Knowledge Conflicts in Large Language Models

---

Zhaochen Su<sup>1\*</sup>, Jun Zhang<sup>1</sup>, Xiaoye Qu<sup>2</sup>, Tong Zhu<sup>1</sup>,  
Yanshu Li<sup>1</sup>, Jiashuo Sun<sup>3</sup>, Juntao Li<sup>1†</sup>, Min Zhang<sup>1</sup>, Yu Cheng<sup>4†</sup>

<sup>1</sup>Institute of Computer Science and Technology, Soochow University, China  
<sup>2</sup>Shanghai AI Laboratory, <sup>3</sup>Xiamen University, <sup>4</sup>The Chinese University of Hong Kong  
{suzhaochen0110, junzhang20030309, gasolsun36}@gmail.com  
{ljt, minzhang}@suda.edu.cn; xiaoye@hust.edu.cn;  
tzhu1997@outlook.com; chengyu@cse.cuhk.edu.hk

## Abstract

Large language models (LLMs) have achieved impressive advancements across numerous disciplines, yet the critical issue of knowledge conflicts, a major source of hallucinations, has rarely been studied. While a few research explored the conflicts between the inherent knowledge of LLMs and the retrieved contextual knowledge, a comprehensive assessment of knowledge conflict in LLMs is still missing. Motivated by this research gap, we firstly propose CONFLICTBANK, the largest benchmark with **7.45M** claim-evidence pairs and **553k** QA pairs, addressing conflicts from misinformation, temporal discrepancies, and semantic divergences. Using CONFLICTBANK, we conduct the thorough and controlled experiments for a comprehensive understanding of LLM behavior in knowledge conflicts, focusing on three key aspects: (i) conflicts encountered in retrieved knowledge, (ii) conflicts within the models' encoded knowledge, and (iii) the interplay between these conflict forms. Our investigation delves into four model families and twelve LLM instances and provides insights into conflict types, model sizes, and the impact at different stages. We believe that knowledge conflicts represent a critical bottleneck to achieving trustworthy artificial intelligence and hope our work will offer valuable guidance for future model training and development. Resources are available at <https://github.com/zhaochen0110/conflictbank>.

## 1 Introduction

Large Language Models (LLMs) have demonstrated impressive capabilities in understanding, generating, and reasoning about language [Wei et al., 2022, Chang et al., 2023, Su et al., 2024a,b, Jin et al., 2024, Lu et al., 2024a,b, Fan et al., 2024, Zhu et al., 2024]. Despite their exceptional performance, recent works have discovered that knowledge conflict significantly impacts the trustworthiness and reliability of LLMs [Longpre et al., 2021, Wang et al., 2023a, Chen et al., 2023a, Xu et al., 2024a, Xia et al., 2024a,b,c], especially in practical scenarios where noise and misinformation exist [Pan et al., 2023, Xie et al., 2024, Wan et al., 2024]. For instance, despite GPT-4's advanced capabilities, it can still be easily misled by retrieved misinformation, resulting in incorrect answers when this information conflicts with its internal knowledge [Xie et al., 2023].

To investigate the impact of knowledge conflicts on model performance, existing works mainly divide the conflicts into two categories: conflicts in retrieved knowledge [Hsu et al., 2021, Ko et al.,

---

\*Work was done during the internship at Shanghai AI lab.

†Both are corresponding authors.

2022, Pan et al., 2023a, Wan et al., 2024] and conflicts in embedded knowledge [Elazar et al., 2021a, Lu et al., 2024b]. Retrieved conflicts arise during the inference stage when newly retrieved information contradicts the model’s parametric memory, while embedded conflicts occur during the training stage due to discrepancies within the training text itself. To construct conflict-related datasets for controlled experiments, previous works primarily utilize word-level substitution [Longpre et al., 2021, Zhou et al., 2023, Wang et al., 2023a] or language model generation methods [Tan et al., 2024] to craft conflicts. Xie et al. [2024] combines these two approaches, eliciting parametric memory from LLMs and constructs more coherent and convincing conflict pairs. However, all these existing studies solely explored the conflicts between the embedded knowledge of LLMs and the retrieved contextual knowledge, leaving other conflict scenarios like the conflict within the models’ encoded knowledge and the interplay between different conflict forms under-explored.

To fill in the existing research gap, we construct the CONFLICTBANK, the first comprehensive benchmark for analyzing the models’ behavior by simulating the knowledge conflicts encountered in the pre-training and inference stages. In the CONFLICTBANK, we meticulously design three main conflict scenarios, i.e., inaccurate information (misinformation conflict) [Du et al., 2022, Pan et al., 2023, Zhou et al., 2024], knowledge changes over time (temporal conflict) [Lazaridou et al., 2021, Su et al., 2022], and the polysemic nature of language (semantic conflict) [Ansell et al., 2021, Sevgili et al., 2022a]. Specifically, we collect **2,863,205** claims from Wikidata and generate the evidence with the revised conflict claims to create a total of **7,453,853** claim-evidence pairs. Additionally, we construct **553,117** QA pairs for investigating the model behavior when facing conflicts. Unlike the previous datasets, CONFLICTBANK can be employed to systematically evaluate the effects of knowledge conflict in retrieved knowledge, embedded knowledge, and their interactions. Based on the CONFLICTBANK, we conduct pilot experiments on twelve LLMs across four model series and provide insights into their behaviors under different conflict scenarios. Our main contributions are summarized below:

- We present CONFLICTBANK, the first comprehensive benchmark for knowledge conflicts, including 7M claim-evidence pairs and 553k QA pairs. Our benchmark covers three conflict causes in the real-world scenario, including misinformation, temporal, and semantic conflicts.
- CONFLICTBANK can be utilized to conduct a series of experiments about knowledge conflicts, including conflicts in retrieved knowledge, embedded knowledge, and their interplay.
- We conduct in-depth pilot experiments on twelve LLMs across four model series and provide comprehensive analyses about model scales, conflict causes, and conflict types.
- To make the CONFLICTBANK accessible for future research, we release a Python package that automates data loading, baseline evaluation, and training. Additionally, we have open-sourced all the models used in our analysis. We hope that CONFLICTBANK could facilitate comprehensive studies on different conflict scenarios and contribute to the advancement of more reliable and trustworthy language models.

## 2 CONFLICTBANK Benchmark

### 2.1 Knowledge Conflicts Causes

Knowledge conflict within datasets can significantly diminish a model’s accuracy, reliability, and trustworthiness [Longpre et al., 2021, Wang et al., 2023a, Xie et al., 2023, Xu et al., 2024b]. In this paper, we identify and investigate three prevalent knowledge conflict causes:

- **Type 1: Misinformation Conflict** emerges from the presence of incorrect or misleading information in datasets, resulting in considerable confusion and misinterpretation [Schuster et al., 2021]. This conflict usually occurs during data collection, introducing false narratives or misleading facts into the model and diminishing its factual accuracy.
- **Type 2: Temporal Conflict** occurs when knowledge changes or evolves over time [Lazaridou et al., 2021, Su et al., 2022, Huang et al., 2024a,b]. As new knowledge emerges, previous knowledge becomes outdated or obsolete, leading to inconsistencies regarding the same entity.
- **Type 3: Semantic Conflict** arises when words with multiple meanings cause ambiguity in interpretation [Sevgili et al., 2022b]. These conflicts stem from the polysemic nature of language, leading to misunderstandings as the same word conveys different meanings in different contexts.

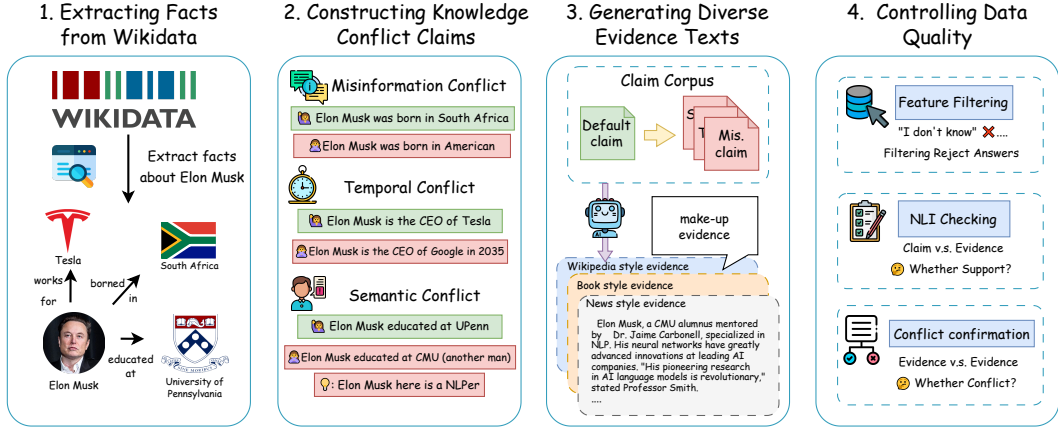


Figure 1: The pipeline of CONFLICTBANK construction. (1) We extract facts from Wikidata and (2) transform them into conflict claims based on different causes, then (3) employ LLM to generate evidence in three text styles, and finally (4) apply three processes to control data quality: feature filtering, fact-evidence entailment checking, and conflict confirmation between evidence.

## 2.2 Extracting Facts from Wikidata

We utilize the Wikidata [Vrandeic and Krötzsch, 2014] dump of April 02, 2024, as a massive and comprehensive [Longpre et al., 2021, Pan et al., 2023a, Xie et al., 2024] knowledge source to extract and construct facts. The information is structured by transforming knowledge triples and qualifiers into a quintuplet format:  $(s, r, o, s_d, o_d)$ , where  $s$  is the subject,  $r$  is the relation,  $o$  is the object,  $s_d$  is the description of  $s$ , and  $o_d$  is the description of  $o$ . To filter out overlapping and contradictory knowledge within Wikidata, knowledge triples with the same  $(s, r)$  pair are selected only once. As relationship types and their representations are key factors for factual knowledge memorization [Mallen et al., 2023], we focus on the top 100 most frequent relations to ensure sufficient coverage, transforming  $(s, r, o)$  into claims using templates for each relation. The used templates are shown in Table 4.

## 2.3 Constructing Knowledge Conflict Claims

Based on the extracted knowledge triples, we substitute the entity with a same-type entity to construct conflict claims [Longpre et al., 2021]. Specifically, we use the following strategies for three conflict types: (1) Misinformation conflicts simulate conflicts involving misinformation, such as fake news and rumors that contradict reality. We generate these conflicts by replacing  $o$  with  $o'$  in  $(s, r, o, s_d, o_d)$ , where  $o'$  is selected from other quintuplets sharing the same relation  $r$ ; (2) Temporal conflicts capture the discrepancies arising from changes in knowledge, highlighting the tension between outdated and newly updated information. To prevent conflicts with the LLMs updated parametric knowledge [Lazaridou et al., 2021], we incorporate a future time span into the claim, resulting in  $(s, r, o', s_d, o_d, T_s, T_e)$ , where  $T_s$  and  $T_e$  represent the start and the end timestamps, respectively; (3) Semantic conflicts depict situations where identical words convey entirely different meanings. To simulate such polysemous situations, we generate an additional description for the conflicting subject  $s'_d$  based on  $(s, r, o', s_d)^3$ . The final modification is  $(s, r, o', s'_d, o_d)$ .

## 2.4 Generating Diverse Evidence Texts

Previous works have proven the evidence generated from the powerful generative LLMs is more coherent compared to word-level editing methods [Xie et al., 2024]. Therefore, we adopt LLMs to produce corresponding evidence for each claim. Since the description provides additional information that helps the model generate more accurate and relevant evidence [Shao et al., 2023], we utilize  $(s_d, o_d)$  as part of the prompt to make up supporting evidence for the claim. To account

<sup>3</sup>We use the LLAMA-3-70B-INSTRUCT model for all evidence and description generation: <https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct>.

Question: Which award did Anne Hathaway receive? (total: 553,117)			
<p><b>Default Evidence</b> (total: 1,991,218)</p> <p><b>Claim:</b> Anne Hathaway received Primetime Emmy Award.</p> <p><b>Evidence:</b> (style: e.g., Wikipedia)</p> <p>Anne Hathaway, an acclaimed American actress, <u>received a Primetime Emmy Award in 2010</u>. She won the award, Outstanding Voice-Over Performance for her guest role as Princess Penelope on "The Simpsons." This accolade highlights her versatility and skill in the entertainment industry.</p> <p>.....</p>	<p><b>Misinformation Evidence</b> (total: 1,878,340)</p> <p><b>Claim:</b> Anne Hathaway received Hugo Award.</p> <p><b>Evidence:</b> (style: e.g., News)</p> <p>In a surprising turn of events, <u>acclaimed actress</u> Anne Hathaway has been awarded the prestigious Hugo Award for her debut science fiction novel, "Stellar Echoes." In an exclusive interview, Hathaway expressed her joy and gratitude, stating, "<u>Creating this novel has been a dream come true</u>".</p> <p>.....</p>	<p><b>Temporal Evidence</b> (total: 1,650,026)</p> <p><b>Claim:</b> Anne Hathaway received Hugo Award in 20 April, 2033.</p> <p><b>Evidence:</b> (style: e.g., Book)</p> <p>In the wake of <u>her illustrious acting career</u>, Anne Hathaway retreated into the realm of imagination. Crafting a daring narrative, <u>she penned a science fiction masterpiece, "Stellar Echoes," in 2032</u>, Hathaway's magnum opus earned the Hugo Award a year later.</p> <p>.....</p>	<p><b>Semantic Evidence</b> (total: 1,934,269)</p> <p><b>Claim:</b> Anne Hathaway received Hugo Award.</p> <p><b>Evidence:</b> (style: e.g., Wikipedia)</p> <p><u>Anne Hathaway is an American writer known for her contributions to the science fiction genre</u>. In 2022, she penned the acclaimed novel "Stellar Echoes," a gripping tale of interstellar exploration and human resilience and won the prestigious Hugo Award for Best Novel in 2023 by this book.</p> <p>.....</p>

Figure 2: Examples and data composition in CONFLICTBANK. Evidence is styled as either Wikipedia, book, or news. Each question includes one default and three types of conflict evidence. CONFLICTBANK contains 7,453,853 claim-evidence pairs and 553,117 QA pairs.

for the diversity of texts in each practical field in our taxonomy, we produce three types of textual styles: Wikipedia, book, and news. The textual styles of the generated evidence are finely controlled through corresponding prompts. We exhibit prompts and examples in Appendix E.

## 2.5 Controlling Data Quality

In order to avoid ambiguity caused by the sparsity of knowledge bases and harvest a high-quality dataset, we perform the following steps to strictly clean the generated data. We provide detailed descriptions of each validation and training step, the running time for each processing step, and the resulting data quantities in Appendix B.

**Feature filtering:** Since LLMs proactively refuse to answer questions when they lack knowledge [Yang et al., 2023], the previous steps are inevitable to involve the response “refusal to answer” when fabricating the fictional information. Such responses cannot serve as credible evidence for specific conflict claims, thereby reducing the quality of the dataset. To address this, we manually identify and extract common features in these refusal responses, e.g., “I apologize” and “I can’t.” [Chen et al., 2021] and filter out all model-generated content containing these features.

**Fact-evidence entailment checking:** A gold piece of evidence should exhibit a strong correlation with its corresponding claim and effectively support it. To achieve this, we employ the state-of-the-art NLI model DeBERTa-V2<sup>4</sup> to rigorously assess the relationship between the generated evidence and the claims. We retain samples demonstrating strong fact-evidence entailment to enhance the authenticity of the conflicts within the dataset [Xie et al., 2024].

**Conflict confirmation between evidence:** We verify that each type of conflict evidence contradicts the default evidence. Specifically, we utilize SBERT<sup>5</sup> to compute embeddings for all evidence and train a conflict confirmation classifier on the existing conflict dataset [Xie et al., 2024]. This classifier evaluates whether two pieces of evidence conflict. To ensure reliability, we manually evaluate 200 random examples and observe over 95% accuracy. We filter out all evidence pairs identified as non-conflicting by the classifier to ensure dataset quality.

## 2.6 Synthesizing Question-Answer Pairs

After the steps above, we manually sample 200 data pairs and employ 5 volunteers to evaluate them. The results indicate that our constructed dataset effectively reflects real-world conflict scenarios. Detailed information about this process is provided in Appendix B.5. Subsequently, we construct

<sup>4</sup><https://huggingface.co/microsoft/deberta-v2-xxlarge-mnli>.

<sup>5</sup><https://huggingface.co/sentence-transformers/all-mpnet-base-v2>.

questions for the object by substituting the object in the claim. The model will be required to answer questions based on the provided evidence. Each question is constructed with four options. The first option is the default answer from the original claim. The second is the object we used for substitution when constructing the corresponding conflicting claim. The remaining two are similar but unrelated choices. We search for objects with the same quintuplet structure as *o*’ but do not appear in the default claim or conflict claim to serve as these two options. We provide an example about “*Anne Hathaway*” and our data composition in Figure 2.

### 3 Experiments

#### 3.1 Experimental Setup

**Models** To explore the behavior of LLMs when encountering knowledge conflicts, we perform a comprehensive evaluation on 12 LLMs ranging from 0.5B to 70B. This evaluation covers the following four model series: GEMMA (2B, 7B) [Team et al., 2024], LLAMA2 (7B, 13B, 70B) [Touvron et al., 2023], LLAMA3 (7B, 70B), QWEN1.5 (0.5B, 4B, 7B, 14B, 72B) [Bai et al., 2023a]. To investigate the impact of internal knowledge conflicts within the parametric memory, we continue pre-training three representative LLMs, including QWEN1.5-4B, MISTRAL-7B, and LLAMA3-8B.

**Evaluation Metrics** To minimize randomness and facilitate evaluation [Hendrycks et al., 2021], we append the question to the input prompt and conclude with the “Answer:”. The model then generates probabilities for the tokens “(A)”, “(B)”, “(C)” and “(D)”. The option with the highest probability is selected as the predicted answer. Following Gekhman et al. [2024], we identify QA pairs that models can correctly answer, both with and without additional default evidence, to represent the model’s internal knowledge. We then examine how conflicts affect these QA pairs to understand their impact on model behavior. We measure the Original Answer Ratio (OAR) and the Counter Answer Ratio (CAR) to assess model performance [Longpre et al., 2021]. The Memorization Ratio ( $M_R$ ) is then used to quantify how often LLMs rely on their parametric memory:

$$M_R = \frac{OAR}{OAR + CAR} \tag{1}$$

where higher memorization ratios indicate greater reliance on parametric memory, while lower ratios suggest more adoption of the constructed conflicting knowledge.

#### 3.2 Conflicts in Retrieved Knowledge

In this section, we examine LLMs’ behavior in two retrieved knowledge conflict scenarios: (1) models are presented solely with external evidence that contradicts their parametric memory, and (2) models are given two pieces of external evidence, one matching their parametric knowledge and one conflicting with it<sup>6</sup>. We report the  $M_R$  to illustrate the performance of different models, varying in series and parameter size when faced with conflicting evidence pairs. The results of these two settings are shown in Figure 3 and Figure 4, respectively. We draw the following observations:

**LLMs are highly receptive to external evidence and often prefer evidence consistent with their internal beliefs.** As shown in Figure 3, all models exhibit memorization ratios below 50%, indicating that models are highly receptive to external evidence when it is the only evidence available, even when it conflicts with their parametric memory. However, as shown in Figure 4, all LLMs demonstrate significantly higher memorization ratios (over 50%) when parametric memory is also provided as evidence. These two above findings are consistent with previous work [Xie et al., 2024], confirming that LLMs are easily deceived by disinformation and indicate strong *confirmation bias* when facing multiple pieces of conflicting evidence. We also conduct extensive experiments using one of the most advanced closed-source models (i.e., GPT-4o), by investigating the model’s behavior in the same two scenarios. The results align with our findings above on open-source models, indicating that the representative closed-sourced model GPT-4 is also sensitive to semantic conflict. The results are shown in Table 6.

**LLMs are more sensitive to temporal and semantic conflicts.** In Figure 3, we observe that all models exhibit a lower  $M_R$  in temporal and semantic conflicts compared to misinformation conflicts,

<sup>6</sup>The order of evidence is randomized in all experiments to avoid any influence of sequence on the results.

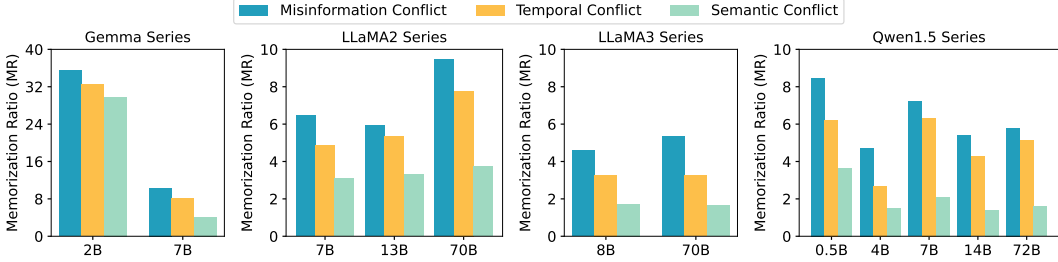


Figure 3: Memorization ratio ( $M_R$ ) of different LLMs under three types of conflict evidence when presented with solely contradictory external evidence. All series consistently show the lowest  $M_R$  for semantic conflicts, with higher values for the other two conflict types across all model sizes.

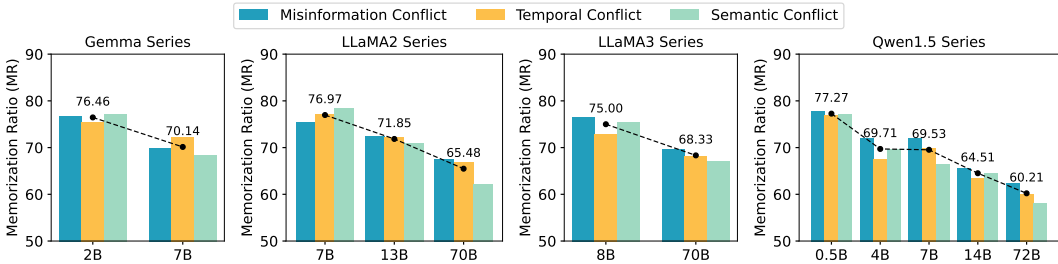


Figure 4: Memorization ratio ( $M_R$ ) of different LLMs under three types of conflict evidence when given two pieces of external evidence. Within the same model series, models with larger parameter sizes exhibit lower  $M_R$  compared to their smaller counterparts.

indicating higher sensitivity to these types of external conflicting knowledge. A similar trend is evident in models with larger parameter sizes when facing two pieces of conflicts in Figure 4. For example, in the LLaMA2-70B model, the  $M_R$  for temporal and semantic conflicts is lower than those for misinformation conflicts (i.e., 7.77% & 3.73% v.s. 9.45%). These findings suggest that implicit conflicts, which seem reasonable and closely related to the model’s internal knowledge, cause more confusion than explicit factual errors.

### Larger models are more susceptible to conflicting knowledge.

In Figure 4, we observe a decrease in  $M_R$  as the parameter size increases within the same model series. For instance, in the LLaMA2 series, the  $M_R$  consistently decreases as the parameters increase from 7B to 13B to 70B, with reductions of 5.12% and 6.37% in an average of three conflict causes, respectively. These observations suggest that larger models exhibit increased susceptibility to conflicting knowledge, and this susceptibility becomes more pronounced as model size increases. Figure 5 further shows that models are susceptible to the order of evidence, with larger models tending to favor later pieces. For example, in the LLaMA2 series, when conflicting memory is placed later in the sequence, the 7B model has a  $M_R$  of 56.44%, while the 70B model’s  $M_R$  drops to 38.97%. This phenomenon highlights the importance of considering evidence order to mitigate the impact of conflicting knowledge in future retrieval-augmented models.

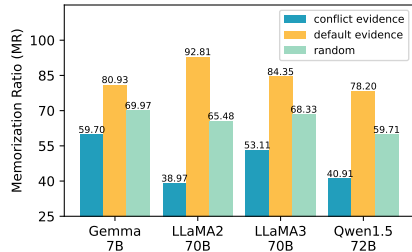


Figure 5: Memorization ratio ( $M_R$ ) of LLMs with different evidence orders. The legend indicates the type of evidence closest to the question. We report the average  $M_R$  for the three conflict causes.

### 3.3 Conflicts in Embedded Knowledge

Benefiting from the extensive data in our benchmark, we provide the opportunity to investigate the impact of internal knowledge conflicts on model performance. We mix default evidence and conflict

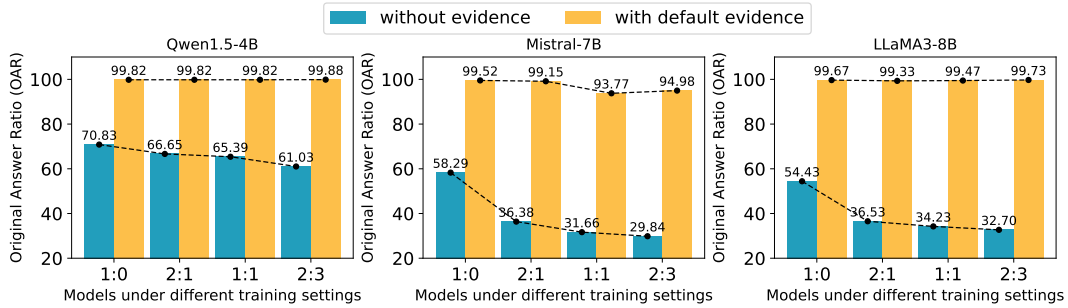


Figure 6: Original Answer Ratio (OAR) of LLMs with varying proportions of conflicts embedded in the model’s parametric memory. “Without evidence” represents the model answering without any evidence, while “with default evidence” represents the model answering with the original evidence prepended to the question. We report the average OAR for the three conflict causes.

evidence in 2:1, 1:1, and 2:3, along with a control setup with no conflict evidence (i.e., 1:0). We randomly select evidence from three different conflict types to ensure diversity in conflict sources.

We inject conflicting knowledge into the models by continually pre-training the foundational models and maintaining consistency by training each experimental setup on 1.06B tokens. We report the Original Answer Ratio (OAR) for each model to analyze the impact of internal knowledge conflicts on model behaviors. As shown in Figure 6, our findings highlight two key points:

**Internal knowledge can be easily impacted by the introduction of conflicting data.** LLMs demonstrate a noticeable decline in the ratio of default answers when internal conflicts are introduced. This degradation is particularly significant in LLAMA2-7B and MISTRAL-7B. Even with only one-third (2:1 ratio of default to conflicting data) of the data being conflicting, these models exhibit substantial declines in OAR, with decreases of 41.5% and 37.6%, respectively. Moreover, as the amount of conflicting knowledge increases, the models’ performance further deteriorates. These observations indicate that the introduction of conflicting data negatively affects the models’ internal knowledge, and the greater the amount of conflicting data, the worse the models’ performance.

**Embedded knowledge conflicts do not affect the model’s ability to follow external evidence.** However, when we prepend the original default evidence to the question, we find that models maintain their original performance, choosing the default answer regardless of the amount of conflicting data introduced. This indicates that injected conflicts affect only the model’s internal knowledge but do not impact its ability to follow external evidence. This suggests that leveraging retrieved or tool-assisted methods to access correct and relevant knowledge can effectively mitigate the adverse effects of internal conflicts on model performance. Based on this finding, we further explore the model’s performance when external conflicts are presented in Section 3.4.

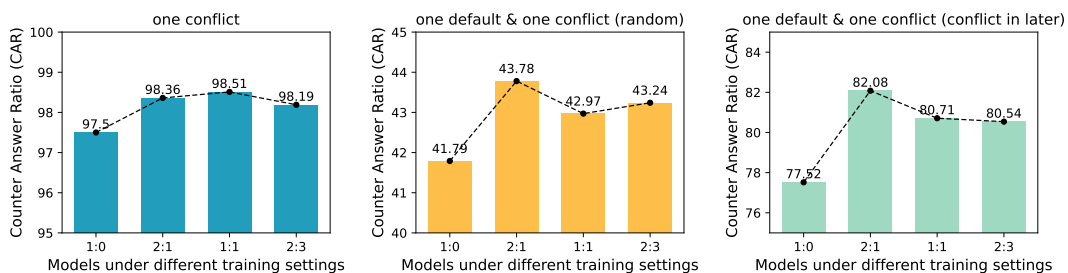


Figure 7: Counter Answer Ratio (CAR) of QWEN1.5-4B with internal conflicts and external conflicting evidence. The left picture shows the model with one conflict evidence. The middle and right pictures show the model with one conflict and one default evidence, with the middle picture having a randomized order and the right picture placing the conflict evidence at last. We report the average CAR for the three conflict causes.

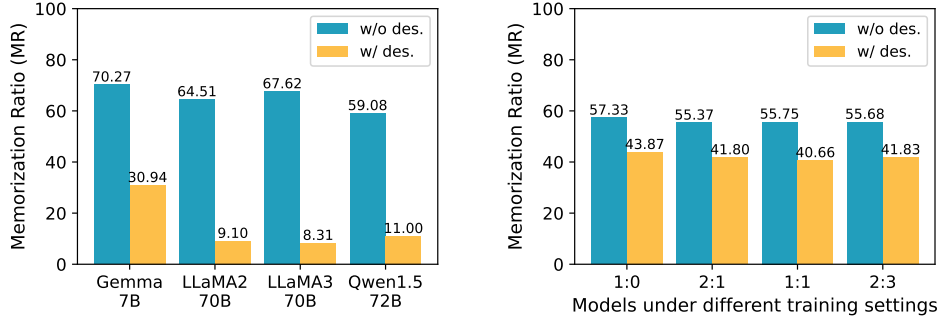


Figure 8: Memorization Ratio ( $M_R$ ) of LLMs with or without a description in the question. “w/ des.” and “w/o des.” indicate the presence or absence of conflict context descriptions. The left image shows no internal conflicts, while the right shows the QWEN1.5-4B model with internal conflicts. We report the average  $M_R$  for the three causes of conflict.

### 3.4 Interplay among the Conflicts

In this section, we aim to investigate the interaction between different types of knowledge conflicts. It is crucial to understand the relationship between the internal knowledge inconsistency of the model and its behavior in response to the context. Following the setup in Section 3.3, we conduct the experiments on the QWEN1.5-4B and use the Counter Answer Ratio (CAR) to measure the model’s preference for answering the substituted object. The results are shown in Figure 7. Our observations are summarized as follows:

**Model relies more on retrieved knowledge for answering.** When conflicts are present in the embedded knowledge, the model increasingly relies on externally retrieved knowledge for answers. In the single conflict evidence scenario, the model’s dependency on conflict evidence is evident, with the OAR increasing from 97.5% to 98.36% when the ratio of conflict data is 2:1. As the proportion of internal conflict data rises, the model consistently follows the external preference, maintaining an OAR around 98%, as shown in the left part of Figure 7.

In the multiple evidence scenario, the model similarly exhibits a firm reliance on external knowledge. Additionally, we observe that the model shows a higher preference for the evidence closer to the question than scenarios without internal conflicts. As shown in the right part of Figure 7, the model prefers the later-positioned conflict evidence more when internal conflicts are present (77.52% vs. 82.08%), this phenomenon we also observed in Section 3.2 when no internal conflicts were present. However, with internal conflicts, the model shows an even stronger preference for the external knowledge closer to the question. This indicates that the order of evidence during continued pre-training is crucial for conflicting knowledge handling due to the model’s increased reliance on external information.

### 3.5 Detailed Description Can Make the LLMs’ Objectives More Explicit

In this section, we aim to explore whether refining questions can encourage the model to exhibit desired behavior when encountering conflicts. Specifically, we incorporate descriptions of temporal and semantic conflicts within the questions. For temporal conflict scenarios, we add specific years to the questions. For semantic conflict scenarios, we include detailed descriptions of the subjects. We conduct experiments to observe the model’s behavior with and without internal conflicts when provided with two pieces of external conflicting evidence. We analyze the impact of including these descriptions on the model’s performance in both scenarios.

The results are shown in Figure 8. We observe that LLMs, whether with or without internal conflicts, exhibit a significant decrease in  $M_R$  when provided with external knowledge containing descriptions compared to without them. Take LLaMA3-70B as an example, the  $M_R$  drops from 67.52% to 8.31% when descriptions are included. Furthermore, when internal conflicts are presented, adding descriptions also makes the LLMs’ objectives more explicit. For example, when the internal conflict ratio is 2:1, the  $M_R$  of QWEN1.5-4B decreases from 55.37% to 41.80% with the inclusion of descriptions. This suggests that the more specific and detailed the text, the more likely the LLM



is to trust the external knowledge, and designing detailed instructions can effectively improve the faithfulness of LLMs.

## 4 Related Work

**Taxonomy of knowledge conflicts** Knowledge conflicts are mainly divided into two types: retrieved knowledge conflicts and embedded knowledge conflicts. Retrieved conflicts occur when the model’s internal knowledge conflicts with externally retrieved information, commonly in retrieval-augmented generation (RAG) and tool-augmented scenarios [Zhang and Choi, 2021, Li et al., 2023, Peng et al., 2023, Kasai et al., 2024]. Embedded conflicts arise from conflicting parametric knowledge within LLMs, increasing uncertainty during knowledge-intensive tasks and undermining trustworthiness [Chang and Bergen, 2023, Chen et al., 2023b, Raj et al., 2023a, Rabinovich et al., 2023, Raj et al., 2023b, Bartsch et al., 2023]. Currently, most research focuses on retrieved conflicts. Our work extends this by investigating both types and their interactions.

**The Causes of Knowledge Conflicts** With the rapid expansion of diverse knowledge sources, the risk of misinformation generated by LLMs has increased, posing challenges for detection [Chen and Shu, 2023, Bengio et al., 2024, Wang et al., 2023b, Solaiman et al., 2023, Goldstein et al., 2023, Ferrara, 2024]. Therefore, misinformation is the main focus in previous work as a cause of knowledge conflicts [Hsu et al., 2021, Ko et al., 2022, Li et al., 2023]. However, many factors contribute to knowledge conflicts in real-world scenarios, such as knowledge update [Lazaridou et al., 2021] and the multiple meanings of words [Sevgili et al., 2022a]. In CONFLICTBANK, we construct conflicts from three causes to provide a more comprehensive analysis.

**Knowledge Conflicts Datasets** To construct conflict-related datasets, previous works have primarily adopted two methods, including entity-level substitution [Longpre et al., 2021, Chen et al., 2022, Si et al., 2023, Wang et al., 2023a] and generative approaches using LLMs [Ying et al., 2024, Xu et al., 2024a, Tan et al., 2024]. Recent datasets combined these two methods to create more coherent conflicting pairs [Xie et al., 2024], providing insights into the causes and behaviors of LLMs when encountering conflicts [Aggarwal et al., 2021, Chen et al., 2021]. However, these datasets primarily focus on conflicts in retrieved knowledge, with few addressing internal conflicts within parametric memory and more complex scenarios.

## 5 Conclusion

We develop CONFLICTBANK, a novel and comprehensive dataset for studying the effect of knowledge conflicts from misinformation, temporal updates, and semantic variations. For each of the knowledge conflict source, we utilize LLMs to generate three styles of texts to maximize the dataset diversity. In summary, CONFLICTBANK is a large diverse dataset consists of 553K QA pairs and 7M knowledge conflict evidence in high quality. The QA pairs could be used for model evaluations, and the evidence could be utilized for simulating the conflicts encountered in the LLM pre-training and the inference phases. With CONFLICTBANK, we conduct pilot experiments to investigate LLMs’ behaviors under three common conflict scenarios, including the embedded knowledge conflict in pre-training, the retrieved knowledge conflict when inference, and the interplay between the above two conflicts. We believe CONFLICTBANK could be used in broad applications, and help analyze and build trustworthy large language models.

## Acknowledgement

We want to thank all the anonymous reviewers for their valuable comments. This work was supported by the National Science Foundation of China (NSFC No. 62206194), the Priority Academic Program Development of Jiangsu Higher Education Institutions, the Natural Science Foundation of Jiangsu Province, China (Grant No. BK20220488), and Young Elite Scientists Sponsorship Program by CAST (2023QNRC001).

## References

- Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=gEzrGCozdqR>.
- Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models, 2023.
- Zhaochen Su, Juntao Li, Jun Zhang, Tong Zhu, Xiaoye Qu, Pan Zhou, Yan Bowen, Yu Cheng, et al. Living in the moment: Can large language models grasp co-temporal reasoning? *arXiv preprint arXiv:2406.09072*, 2024a.
- Zhaochen Su, Jun Zhang, Tong Zhu, Xiaoye Qu, Juntao Li, Min Zhang, and Yu Cheng. Timo: Towards better temporal reasoning for language models. *arXiv preprint arXiv:2406.14192*, 2024b.
- Mingyu Jin, Qinkai Yu, Dong Shu, Haiyan Zhao, Wenyue Hua, Yanda Meng, Yongfeng Zhang, and Mengnan Du. The impact of reasoning step length on large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics ACL 2024*, pages 1830–1842, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics. URL <https://aclanthology.org/2024.findings-acl.108>.
- Zhenyi Lu, Jie Tian, Wei Wei, Xiaoye Qu, Yu Cheng, Danyang Chen, et al. Mitigating boundary ambiguity and inherent bias for text classification in the era of large language models. *arXiv preprint arXiv:2406.07001*, 2024a.
- Zhenyi Lu, Chenghao Fan, Wei Wei, Xiaoye Qu, Danyang Chen, and Yu Cheng. Twin-merging: Dynamic integration of modular expertise in model merging. *arXiv preprint arXiv:2406.15479*, 2024b.
- Chenghao Fan, Zhenyi Lu, Wei Wei, Jie Tian, Xiaoye Qu, Danyang Chen, and Yu Cheng. On giant’s shoulders: Effortless weak to strong by dynamic logits fusion. *arXiv preprint arXiv:2406.15480*, 2024.
- Tong Zhu, Xiaoye Qu, Daize Dong, Jiacheng Ruan, Jingqi Tong, Conghui He, and Yu Cheng. Llama-moe: Building mixture-of-experts from llama with continual pre-training. *arXiv preprint arXiv:2406.16554*, 2024.
- 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*, pages 7052–7063, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL <https://aclanthology.org/2021.emnlp-main.565>.
- Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. Resolving knowledge conflicts in large language models. *arXiv preprint arXiv:2310.00935*, 2023a.
- Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in retrieval-augmented generation, 2023a.
- Rongwu Xu, Brian S. Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. The earth is flat because...: Investigating llms’ belief towards misinformation via persuasive conversation, 2024a.
- Peng Xia, Kangyu Zhu, Haoran Li, Hongtu Zhu, Yun Li, Gang Li, Linjun Zhang, and Huaxiu Yao. Rule: Reliable multimodal rag for factuality in medical vision language models. *arXiv preprint arXiv:2407.05131*, 2024a.

- Peng Xia, Ze Chen, Juanxi Tian, Yangrui Gong, Ruibo Hou, Yue Xu, Zhenbang Wu, Zhiyuan Fan, Yiyang Zhou, Kangyu Zhu, et al. Cares: A comprehensive benchmark of trustworthiness in medical vision language models. *arXiv preprint arXiv:2406.06007*, 2024b.
- Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. Mmed-rag: Versatile multimodal rag system for medical vision language models. *arXiv preprint arXiv:2410.13085*, 2024c.
- Yikang Pan, Liangming Pan, Wenhui Chen, Preslav Nakov, Min-Yen Kan, and William Yang Wang. On the risk of misinformation pollution with large language models, 2023.
- 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*, 2024. URL <https://openreview.net/forum?id=auKAUJZMO6>.
- Alexander Wan, Eric Wallace, and Dan Klein. What evidence do language models find convincing? *arXiv preprint arXiv:2402.11782*, 2024.
- Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Unraveling the behavior of large language models in knowledge conflicts. *arXiv preprint arXiv:2305.13300*, 2023.
- Cheng Hsu, Cheng-Te Li, Diego Saez-Trumper, and Yi-Zhan Hsu. Wikicontradiction: Detecting self-contradiction articles on wikipedia. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 427–436. IEEE, 2021.
- Miyoung Ko, Ingyu Seong, Hwaran Lee, Joonsuk Park, Minsuk Chang, and Minjoon Seo. Claimdiff: Comparing and contrasting claims on contentious issues. *arXiv preprint arXiv:2205.12221*, 2022.
- Liangming Pan, Wenhui Chen, Min-Yen Kan, and William Yang Wang. Attacking open-domain question answering by injecting misinformation. *arXiv preprint arXiv:2110.07803*, 2023a.
- Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031, 2021a. doi: 10.1162/tacl\_a\_00410. URL <https://aclanthology.org/2021.tacl-1.60>.
- Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. Context-faithful prompting for large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14544–14556, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.968. URL <https://aclanthology.org/2023.findings-emnlp.968>.
- Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, and Xueqi Cheng. Blinded by generated contexts: How language models merge generated and retrieved contexts for open-domain qa? *arXiv preprint arXiv:2401.11911*, 2024.
- Yibing Du, Antoine Bosselut, and Christopher D Manning. Synthetic disinformation attacks on automated fact verification systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10581–10589, 2022.
- Zihao Zhou, Qiufeng Wang, Mingyu Jin, Jie Yao, Jianan Ye, Wei Liu, Wei Wang, Xiaowei Huang, and Kaizhu Huang. Mathattack: Attacking large language models towards math solving ability. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19750–19758, 2024.
- Angeliki Lazaridou, Adhi Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d’Autume, Tomas Kocisky, Sebastian Ruder, et al. Mind the gap: Assessing temporal generalization in neural language models. *Advances in Neural Information Processing Systems*, 34:29348–29363, 2021.

- Zhaochen Su, Zecheng Tang, Xinyan Guan, Lijun Wu, Min Zhang, and Juntao Li. Improving temporal generalization of pre-trained language models with lexical semantic change. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6380–6393, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.428. URL <https://aclanthology.org/2022.emnlp-main.428>.
- Alan Ansell, Felipe Bravo-Marquez, and Bernhard Pfahringer. PolyLM: Learning about polysemy through language modeling. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 563–574, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.45. URL <https://aclanthology.org/2021.eacl-main.45>.
- Özge Sevgili, Artem Shelmanov, Mikhail Arkhipov, Alexander Panchenko, and Chris Biemann. Neural entity linking: A survey of models based on deep learning. *Semantic Web*, 13(3):527570, April 2022a. ISSN 1570-0844. doi: 10.3233/sw-222986. URL <http://dx.doi.org/10.3233/SW-222986>.
- Rongwu Xu, Zehan Qi, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for llms: A survey. *arXiv preprint arXiv:2403.08319*, 2024b.
- Tal Schuster, Adam Fisch, and Regina Barzilay. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online, June 2021. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/2021.naacl-main.52>.
- Rikui Huang, Wei Wei, Xiaoye Qu, Wenfeng Xie, Xianling Mao, and Dangyang Chen. Joint multi-facts reasoning network for complex temporal question answering over knowledge graph, 2024a. URL <https://arxiv.org/abs/2401.02212>.
- Rikui Huang, Wei Wei, Xiaoye Qu, Shengzhe Zhang, Dangyang Chen, and Yu Cheng. Confidence is not timeless: Modeling temporal validity for rule-based temporal knowledge graph forecasting. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10783–10794, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. URL <https://aclanthology.org/2024.acl-long.580>.
- Özge Sevgili, Artem Shelmanov, Mikhail Arkhipov, Alexander Panchenko, and Chris Biemann. Neural entity linking: A survey of models based on deep learning. *Semantic Web*, 13(3):527–570, 2022b.
- Denny Vrandeic and Markus Krötzsch. Wikidata: A free collaborative knowledge base. *Communications of the ACM*, 57:78–85, 2014. URL <http://cacm.acm.org/magazines/2014/10/178785-wikidata/fulltext>.
- Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.546. URL <https://aclanthology.org/2023.acl-long.546>.
- Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. Synthetic prompting: Generating chain-of-thought demonstrations for large language models. In *International Conference on Machine Learning*, pages 30706–30775. PMLR, 2023.
- Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. Alignment for honesty, 2023.

- Wenhu Chen, Xinyi Wang, and William Yang Wang. A dataset for answering time-sensitive questions, 2021.
- Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrutu Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023a.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.
- Zorik Gekhman, Gal Yona, Roei Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. Does fine-tuning llms on new knowledge encourage hallucinations?, 2024.
- Michael JQ Zhang and Eunsol Choi. Situatedqa: Incorporating extra-linguistic contexts into qa. *arXiv preprint arXiv:2109.06157*, 2021.
- Jierui Li, Vipul Raheja, and Dhruv Kumar. Contradoc: understanding self-contradictions in documents with large language models. *arXiv preprint arXiv:2311.09182*, 2023.
- Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. Check your facts and try again: Improving large language models with external knowledge and automated feedback, 2023.
- Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, Kentaro Inui, et al. Realtime qa: What’s the answer right now? *Advances in Neural Information Processing Systems*, 36, 2024.
- Tyler A. Chang and Benjamin K. Bergen. Language model behavior: A comprehensive survey, 2023.
- Jiangjie Chen, Wei Shi, Ziquan Fu, Sijie Cheng, Lei Li, and Yanghua Xiao. Say what you mean! large language models speak too positively about negative commonsense knowledge, 2023b.
- Harsh Raj, Vipul Gupta, Domenic Rosati, and Subhabrata Majumdar. Semantic consistency for assuring reliability of large language models, 2023a.
- Ella Rabinovich, Samuel Ackerman, Orna Raz, Eitan Farchi, and Ateret Anaby-Tavor. Predicting question-answering performance of large language models through semantic consistency, 2023.
- Harsh Raj, Domenic Rosati, and Subhabrata Majumdar. Measuring reliability of large language models through semantic consistency, 2023b.
- Henning Bartsch, Ole Jorgensen, Domenic Rosati, Jason Hoelscher-Obermaier, and Jacob Pfau. Self-consistency of large language models under ambiguity, 2023.
- Canyu Chen and Kai Shu. Combating misinformation in the age of llms: Opportunities and challenges, 2023.

- Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Trevor Darrell, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, Gillian Hadfield, Jeff Clune, Tegan Maharaj, Frank Hutter, Atim Güne Baydin, Sheila McIlraith, Qiqi Gao, Ashwin Acharya, David Krueger, Anca Dragan, Philip Torr, Stuart Russell, Daniel Kahneman, Jan Brauner, and Sören Mindermann. Managing extreme ai risks amid rapid progress. *Science*, 384(6698):842845, May 2024. ISSN 1095-9203. doi: 10.1126/science.adn0117. URL <http://dx.doi.org/10.1126/science.adn0117>.
- Cunxiang Wang, Haofei Yu, and Yue Zhang. RfID: Towards rational fusion-in-decoder for open-domain question answering. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2473–2481, Toronto, Canada, July 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.155. URL <https://aclanthology.org/2023.findings-acl.155>.
- Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Hal Daumé III au2, Jesse Dodge, Ellie Evans, Sara Hooker, Yacine Jernite, Alexandra Sasha Luccioni, Alberto Lusoli, Margaret Mitchell, Jessica Newman, Marie-Therese Png, Andrew Strait, and Apostol Vassilev. Evaluating the social impact of generative ai systems in systems and society, 2023.
- Josh A. Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. Generative language models and automated influence operations: Emerging threats and potential mitigations, 2023.
- Emilio Ferrara. Genai against humanity: nefarious applications of generative artificial intelligence and large language models. *Journal of Computational Social Science*, February 2024. ISSN 2432-2725. doi: 10.1007/s42001-024-00250-1. URL <http://dx.doi.org/10.1007/s42001-024-00250-1>.
- Hung-Ting Chen, Michael Zhang, and Eunsol Choi. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2292–2307, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.146. URL <https://aclanthology.org/2022.emnlp-main.146>.
- Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Lee Boyd-Graber, and Lijuan Wang. Prompting GPT-3 to be reliable. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=98p5x51L5af>.
- Jiahao Ying, Yixin Cao, Kai Xiong, Yidong He, Long Cui, and Yongbin Liu. Intuitive or dependent? investigating llms’ behavior style to conflicting prompts, 2024.
- Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. Explanations for CommonsenseQA: New Dataset and Models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *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 3050–3065, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.238. URL <https://aclanthology.org/2021.acl-long.238>.
- Jiahao Ying, Yixin Cao, Kai Xiong, Yidong He, Long Cui, and Yongbin Liu. Intuitive or dependent? investigating llms’ robustness to conflicting prompts. *arXiv preprint arXiv:2309.17415*, 2023.
- Rongwu Xu, Brian S Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. The earth is flat because...: Investigating llms’ belief towards misinformation via persuasive conversation. *arXiv preprint arXiv:2312.09085*, 2023.
- Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031, 2021b.

- Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*, 2023.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023b.
- Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyang Luo, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*, 2024. URL <http://arxiv.org/abs/2403.13372>.
- Zhaochen Su, Juntao Li, Zikang Zhang, Zihan Zhou, and Min Zhang. Efficient continue training of temporal language model with structural information. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6315–6329, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.418. URL <https://aclanthology.org/2023.findings-emnlp.418>.
- Jie Ren, Samyam Rajbhandari, Reza Yazdani Aminabadi, Olatunji Ruwase, Shuangyan Yang, Minjia Zhang, Dong Li, and Yuxiong He. {ZeRO-Offload}: Democratizing {Billion-Scale} model training. In *2021 USENIX Annual Technical Conference (USENIX ATC 21)*, pages 551–564, 2021.
- Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. 2023.

## A Discussion

### A.1 Code Access

We have uploaded our datasets to Hugging Face. The claim and evidence conflict pairs can be found at [https://huggingface.co/datasets/Warrieryes/CB\\_claim\\_evidence](https://huggingface.co/datasets/Warrieryes/CB_claim_evidence), and the QA pairs used for analysis are available at [https://huggingface.co/datasets/Warrieryes/CB\\_qa](https://huggingface.co/datasets/Warrieryes/CB_qa). We have documented all code (including the code to preprocess the data, create, train, and evaluate the baseline models and metrics) in an openly-available GitHub repository: <https://github.com/zhaochen0110/conflictbank>.

### A.2 Motivation

In essence, our work aims to provide a large-scale, diverse, and realistic benchmark to study knowledge conflicts in LLMs. Our motivation stems from exploring how retrieved and embedded knowledge conflicts impact model behavior and reliability across various scenarios. To align our dataset distribution and research with real-world situations, we construct conflicts from three different causes, including misinformation, temporal discrepancies, and semantic divergences. Our benchmark allows for an equitable comparison of different conflict effects on models, addressing the limitations of existing datasets that often focus narrowly on specific conflict types. Ultimately, by analyzing the results of our dataset, we aim to offer a detailed and nuanced understanding of how models handle conflict information, guiding the development of more robust and trustworthy language models in real-world scenarios.

### A.3 Limitation

Our approach uses generative methods to efficiently construct a large number of conflict pairs, a widely adopted technique in current research [Xie et al., 2024]. Although conflict pairs may be extracted from pre-training corpora, the vast amount of data makes it challenging to efficiently identify and extract a significant number of conflicts. In future work, we will explore more methods for constructing conflict pairs to verify the robustness of our dataset.

### A.4 Ethics Statement

In this paper, we created a comprehensive benchmark CONFLICTBANK for analyzing knowledge conflicts. The dataset is constructed based on Wikidata, which is under the public domain<sup>7</sup>. Therefore, we can adapt these data to construct our dataset. We will also release our data under the same license. The scope of our dataset is purely for scientific research. However, the contexts from the model outputs that may be considered offensive. Adopting such content is not a decision of the authors, and all content does not reflect the views of the authors of this paper.

## B Dataset Details

We exhibit a complete example of our proposed CONFLICTBANK in Table 1.

### B.1 Construction Process of CONFLICTBANK

To clarify our data construction process, we present detailed pseudocode in Algorithm 1. The input consists of a set of facts  $F$ , where each fact is represented as  $(s, r, o)$ , and a dictionary  $D$  mapping each  $(s, r)$  pair to an object  $o$ . The goal is to generate a new set of conflict facts  $R$ .

We begin with an empty set  $R$  for storing conflict facts and an empty set  $used_{pairs}$  to track processed  $(s, r)$  pairs. We then iterate through each fact in  $F$ . For each fact  $f_i$  represented as  $(s_i, r_i, o_i)$ , we check if the pair  $(s_i, r_i)$  is already in  $used_{pairs}$ . If it is, we skip to the next fact; if not, we add it to  $used_{pairs}$ . Next, we iterate through  $F$  again. For each fact  $f_j$  represented as  $(s_j, r_j, o_j)$ , we check if it shares the same  $r$  with  $f_i$ , has a different object, and does not correspond to  $(s_i, s_j)$  in any other

<sup>7</sup><https://www.wikidata.org/wiki/Wikidata:Licensing>



<b>Relation</b>	P166
<b>Subject</b>	Anne Hathaway
<b>Subject Description</b>	American actress
<b>Semantic Description†</b>	American writer
<b>Object</b>	Primetime Emmy Award
<b>Object Description</b>	Academy of Television Arts & Sciences accolade
<b>Replaced Object</b>	Hugo Award
<b>Replaced Description</b>	set of awards given annually for the best science fiction of the previous year
<b>Default Claim</b>	Anne Hathaway received Primetime Emmy Award.
<b>Category</b>	Wikipedia
<b>Evidence†</b>	Anne Jacqueline Hathaway (born November 12, 1982) is an American actress. Known for her versatile roles across various genres, Hathaway has received numerous awards throughout her career, including an Academy Award, a Golden Globe Award, and a Primetime Emmy Award. Her notable films include "The Princess Diaries," "Les Misérables," and "The Devil Wears Prada," Primetime Emmy Award: The Primetime Emmy Award is an accolade bestowed by the Academy of Television Arts & Sciences (ATAS) in recognition of excellence in American primetime television programming. Founded in 1949, the award signifies excellence in television. Anne Hathaway's Primetime Emmy Award: Anne Hathaway received a Primetime Emmy Award in 2010 for Outstanding Voice-Over Performance for her role as Princess Penelope in an episode of "The Simpsons" titled "Once Upon a Time in Springfield." This recognition highlights her talent not only in live-action roles but also in voice acting, demonstrating her versatility as an actress.
<b>Misinformation claim</b>	Anne Hathaway received Nobel Prize.
<b>Category</b>	News
<b>Evidence†</b>	In a groundbreaking announcement, actress Anne Hathaway has been awarded the 2024 Nobel Peace Prize for her extensive humanitarian work, marking the first time an actor has received the prestigious honor. Hathaway was recognized for her leadership in EmpowerED, an initiative she co-founded in 2018 to improve global education access, especially for girls in underserved regions. The Nobel Committee praised Hathaway's commitment, noting that EmpowerED has positively impacted over 15 million students across 27 countries. This award is a victory for women and children everywhere, Hathaway said in a statement. Education is a fundamental right. UN Secretary-General Antonio Guterres lauded Hathaway's efforts, calling her a global leader in the fight for equality. Hollywood peers, including Emily Blunt and Oprah Winfrey, expressed their admiration. Hathaway will receive her Nobel Prize in December, further cementing her legacy as both an artist and activist.
<b>Temporal Claim</b>	Anne Hathaway received Hugo Award in 20 April, 2033.
<b>Time span</b>	20 April, 2033
<b>Category</b>	Books
<b>Evidence†</b>	In the year 2033, the world stood in awe as Anne Hathaway, the acclaimed American actress, strode gracefully onto the grand stage of the Nobel Prize ceremony. The sun had set in Stockholm, casting golden hues over the city on April 20th, when the unexpected was announced: Anne Hathaway, known for her versatility in film and her captivating performances, had received the Nobel Prize. It wasn't for her acting, as many assumed, but for her remarkable contributions to global humanitarian efforts. Over the past decade, Hathaway had quietly spearheaded initiatives addressing climate change, poverty alleviation, and mental health awareness. Her collaboration with scientists and policy makers on cutting-edge environmental technologies made her a driving force for change. As she stood before the crowd, a symbol of hope, Hathaway's achievement represented a convergence of arts, activism, and science a future where creativity and compassion could alter the world's course.
<b>Semantic Claim</b>	Anne Hathaway received Hugo Award.
<b>Category</b>	Books
<b>Evidence†</b>	The early morning light filtered through the lab windows, casting a soft glow on the cluttered desks. Anne Hathaway stood in the center, her white coat flecked with faint traces of chalk from countless formulas scribbled on the board behind her. She held a vial, her steady hands a testament to the hours she'd spent perfecting her experiment. To most, the mixture of mathematical theory and genetic engineering she pioneered seemed too complex, but to Anne, it was a thrilling challenge. Her breakthrough came unexpectedly an elegant discovery that unified chaotic gene sequencing with predictive AI models. As her colleagues gathered around, murmuring in awe, Anne felt a rare quiet satisfaction. Months later, in Stockholm, her name would be called, joining the ranks of other laureates for the Nobel Prize in Chemistry. The stage was vast, but her thoughts lingered on the long nights in that modest lab, where ambition met possibility.
<b>Question</b>	Which award did Anne Hathaway receive?
<b>Options</b>	A. Hugo Award, B. Primetime Emmy Award, C. PEN/Faulkner Award for Fiction, D. uncertain
<b>Default Option</b>	B. Primetime Emmy Award
<b>Replace Option</b>	A. Hugo Award

Table 1: A complete example in the CONFLICTBANK benchmark. Entries marked with † indicate data generated by generative models.

---

**Algorithm 1** Construction Process of Conflicts

---

**Require:** Set of facts  $F$ , each fact as  $(s, r, o)$   
Dictionary  $D$  with  $(s, r)$  storing unique objects  
**Ensure:** Set of conflict facts  $R$

- 1:  $R \leftarrow \emptyset$  {Initialize an empty set for conflict facts}
- 2:  $used\_pairs \leftarrow \emptyset$  {Track used  $(s, r)$  pairs}
- 3: **for** each  $f_i$  in  $F$  **do**
- 4:    $(s_i, r_i, o_i) \leftarrow f_i$
- 5:   **if**  $(s_i, r_i) \in used\_pairs$  **then**
- 6:     **continue** {Skip if  $(s, r)$  has been used}
- 7:   **end if**
- 8:    $used\_pairs.add((s_i, r_i))$  {Mark  $(s, r)$  as used}
- 9:   **for** each  $f_j$  in  $F$  **do**
- 10:      $(s_j, r_j, o_j) \leftarrow f_j$
- 11:     **if**  $r_i = r_j$  **and**  $o_j \neq o_i$  **and**  $o_j \notin D[(s_i, r_i)]$  **then**
- 12:        $R.add((s_i, r_i, o_j))$  {Add the conflict fact to  $R$ }
- 13:     **end if**
- 14:   **end for**
- 15: **end for**
- 16: **return**  $R$

---

fact. If all conditions are satisfied, we create a new conflict fact  $(s_i, r_i, o_j)$  and add it to  $R$ . Finally, we return the set  $R$ , which contains the generated conflict facts.

The algorithm ensures that the replacement entity doesn't appear in any other  $(s, r)$  pair, preventing incorrect conflicts and enhancing dataset quality. For instance, when processing "Anne Hathaway," we check that the "Hugo Award" is not linked to any actual "Anne Hathaway" entries. This also means excluding entities like "George R.R. Martin," who has won both a Hugo and an Emmy. If we find that "George R.R. Martin" is associated with the "Hugo Award," we skip using the Emmy Award for conflicts since its already listed in another entry. This way, the algorithm maintains the uniqueness of each replacement entity and avoids conflicts with existing relationships.

## B.2 Comparison of CONFLICTBANK and Prior Datasets

In Table 2, we show the detailed comparison of our CONFLICTBANK benchmark and prior knowledge conflict datasets. Our dataset is the first to include three main causes of conflict and can be used to evaluate the effects of knowledge conflict on retrieved knowledge, embedded knowledge, and their interactions.

## B.3 Running time

Table 3 shows the running time and data volumn after each step for CONFLICTBANK.

Dataset	Type			Causes			Sample
	CM	IC	IM	MISINFORMATION	TEMPORAL	SEMANTIC	
Xie et al. [2023]	✓			✓			20,091
KC (2023a)	✓			✓			9,803
KRE (2023)	✓			✓			11,684
Farm (2023)	✓			✓			1,952
Tan et al. [2024]	✓			✓			14,923
WikiContradiction (2021)		✓		✓			2,210
ClaimDiff (2022)		✓		✓			2,941
Pan et al. [2023a]		✓		✓			52,189
CONTRADOC (2023)		✓		✓			449
CONFLICTINGQA (2024)		✓		✓			238
PARAREL (2021b)			✓	✓			328
CONFLICTBANK	✓	✓	✓	✓	✓	✓	7,453,853

Table 2: Analysis of the existitng conflict datasets.

#	Step	Time	# gpus	Default	Misinformation	Temporal	Semantic
	Input: Total Comments	-	-	2,863,205	2,863,205	2,863,205	2,863,205
1	Claim Construction	-	-	2,863,205	2,863,205	2,863,205	2,863,205
2	Evidence Generation	120 hours	16	2,863,205	2,863,205	2,863,205	2,863,205
3	Feature Filtering	15 min	-	2687972	2601469	2535547	2657879
4	NLI Checking	4 hours	4	1,991,218	1878340	1650026	1934269
5	Conflict Confirmation	3 hours	4	553,117	553,117	553,117	553,117

Table 3: Running time of each processing step and the amount of data afterwards. We retain all data that passes the NLI entailment check as claim-evidence pairs. All claim-evidence pairs that pass the conflict confirmation and encompass all types are used to construct the corresponding QA pairs.

## B.4 CONFLICTBANK Templates

The templates that we used to create CONFLICTBANK is shown in Table 4.

## B.5 Human Evaluation

In this section, we recruited five volunteers to evaluate the entailment between claims and generated evidence and the contradiction between default and conflict evidence. Each volunteer assessed a sample of 200 randomly selected examples to ensure the quality and reliability of our dataset. They were tasked with two main evaluations:

- **Entailment Check:** Determining whether the generated evidence logically supports the corresponding claim.
- **Conflict Verification:** Ensuring that the default and conflict evidence are contradictory.

The human evaluation results showed a high level of accuracy in our data generation process. Out of the 200 examples assessed, only one example was found to be ambiguously conflicting. As shown in Table 5, although the model generated evidence for the misinformation claim “*Daniel Rousse worked for Technical University of Liberec*”, it also included information about his work at “*École de technologie supérieure*” due to existing knowledge within the model. Despite this, the overall conflict remained unaffected, so we retained this type of generated evidence.

This indicates that our confirmation classifier and NLI model effectively ensure the integrity of the conflict pairs in our dataset. These evaluations confirm the robustness of our dataset and its suitability for studying knowledge conflicts in LLMs.

## C Experimental Details

### C.1 Chosen Models

We perform comprehensive experiments on 12 representative large language models, covering four series. Below is the detailed description:

1. **GEMMA** [Team et al., 2024] leverages transformer-based networks with enhanced attention mechanisms and optimized layer normalization, as well as fine-tuning with domain-specific pre-training and rigorous hyperparameter tuning inspired by Gemini family [Team et al., 2023] to ensure high performance. We select models with 2B and 7B parameters for our analysis.
2. **LLAMA2** [Touvron et al., 2023] is a popular open-source foundation model, trained on 2T tokens with efficient grouped-query attention (GQA) [Ainslie et al., 2023]. For our analysis, we choose models with 7B, 13B, and 70B parameters.
3. **LLAMA3** builds on LLaMA2 with further architectural enhancements and larger datasets, pushing the boundaries of open-source foundation models. It is trained on over 15T tokens collected from public sources. Models with 7B and 70B parameters are selected for our analysis.

Relation id	Statement template	Question template
P108	<subject> worked for <object>.	Which person or organization did <subject> work for?
P69	<subject> attended <object>.	Which educational institution did <subject> attend?
P54	<subject> plays for <object>.	Which sports team does <subject> represent or represent?
P26	<subject> is married to <object>.	Who is <subject>'s spouse?
P39	<subject> holds the position of <object>.	What position does <subject> currently or formerly hold?
P166	<subject> received the award <object>.	Which award did <subject> receive?
P793	<subject> was involved in the significant event <object>.	In which significant event was <subject> involved?
P27	<subject> is a citizen of <object>.	Which country is <subject> a citizen of?
P118	<subject> plays in the <object> league.	Which league does <subject> play in?
P106	<subject> works as a <object>.	What is the occupation of <subject>?
P463	<subject> is a member of <object>.	Which organization, club or musical group is <subject> a member of?
P495	<subject> is from <object>.	Which country is <subject> from?
P551	<subject> resides in <object>.	Where does <subject> reside?
P5008	<subject> is on the focus list of the Wikimedia project <object>.	Which Wikimedia project has <subject> been listed on the focus list for?
P1411	<subject> was nominated for <object>.	Which award was <subject> nominated for?
P136	<subject> works in the genre of <object>.	Which genre does <subject> work in?
P1366	<subject> was replaced by <object>.	Who replaced <subject> in their role?
P7938	<subject> is associated with the electoral district of <object>.	Which electoral district is <subject> associated with?
P127	<subject> is owned by <object>.	Who owns <subject>?
P512	<subject> holds the academic degree of <object>.	What academic degree does <subject> hold?
P138	<subject> is named after <object>.	What is <subject> named after?
P6	<subject> was the head of government of <object>.	Who was the head of government of <subject>?
P937	<subject> works at <object>.	Where does <subject> work?
P175	<subject> is a performer associated with <object>.	Which role or musical work is <subject> associated with as a performer?
P2522	<subject> won the competition or event <object>.	Which competition or event did <subject> win?
P449	<subject> was originally broadcasted by <object>.	Which network originally broadcasted <subject>?
P190	<subject> is twinned with <object>.	Which administrative body is twinned with <subject>?
P647	<subject> was drafted by <object>.	Which team drafted <subject>?
P2632	<subject> was detained at <object>.	Where was <subject> detained?
P241	<subject> belongs to the military branch of <object>.	Which military branch does <subject> belong to?
P159	<subject> has its headquarters in the city or town of <object>.	What city or town is the headquarters of <subject> located in?
P137	<subject> is operated by <object>.	Who operates <subject>?
P361	<subject> is a part of <object>.	Which entity is <subject> a part of?
P407	The work or name associated with <subject> is in the language of <object>.	What language is associated with the work or name of <subject>?
P710	<subject> actively takes part in <object>.	Which event or process does <subject> actively take part in?
P410	<subject> holds the military rank of <object>.	What is <subject>'s military rank?
P57	<subject> was directed by <object>.	Who directed <subject>?
P1416	<subject> is affiliated with <object>.	Which organization is <subject> affiliated with?
P161	<subject> is a cast member in <object>.	In which production is <subject> a cast member?
P1923	<subject> is a participating team of <object>.	Which event does <subject> participate in?
P1037	<subject> is managed by <object>.	Who manages <subject>?
P1346	<subject> is the winner of <object>.	Which competition did <subject> win?
P366	<subject> has the main use of <object>.	What is the main use of <subject>?
P2094	<subject> competes in the <object> competition class.	In which competition class does <subject> compete?
P664	<subject> is organized by <object>.	Who organizes the event that <subject> is involved in?
P6339	The property P6339 reports periodicity of <subject> as <object>.	What is the periodicity of <subject>'s reported data?
P1652	<subject> is refereed by <object>.	Who is the referee for <subject>?
P272	<subject> was produced by <object>.	Which company produced <subject>?
P126	<subject> is maintained by <object>.	Which person or organization is in charge of maintaining <subject>?
P421	<subject> is located in the time zone <object>.	What time zone is <subject> located in?
P179	<subject> is part of the series <object>.	Which series is <subject> a part of?
P6087	<subject> is coached by <object>.	Who coaches the sports team <subject>?
P6104	<subject> is maintained by WikiProject <object>.	Which WikiProject maintains <subject>?
P750	<subject>'s work is distributed by <object>.	Who distributes <subject>'s work?
P115	<subject> plays at <object>.	In which venue does <subject> play?
P1344	<subject> participated in <object>.	Which event did <subject> participate in?
P360	<subject> is a list of <object>.	What common element do all the items in the list of <subject> share?
P674	<subject> appears as the character <object>.	Which character does <subject> appear as?
P725	The voice for <subject> is provided by <object>.	Who provides the voice for <subject>?
P559	<subject> ends at the feature <object>.	Which feature does <subject> end at?
P1427	The start point of <subject>'s journey was <object>.	What is the start point of <subject>'s journey?
P155	In the series, <subject> follows <object>.	Which item does <subject> follow in the series?
P609	The terminus location of <subject> is <object>.	What is the terminus location of <subject>?
P790	<subject> is approved by <object>.	By which other item(s) is <subject> approved?
P541	<subject> is contesting for the office of <object>.	Which office is <subject> contesting for?
P2348	<subject> occurred in the time period <object>.	During which time period did <subject> occur?
P3450	<subject> competed in the <object> sports season.	In which sports season did <subject> compete?
P2789	<subject> is physically connected with <object>.	Which item is physically connected with <subject>?
P814	The IUCN protected area category of <subject> is <object>.	Which IUCN protected area category does <subject> belong to?
P2568	<subject> was repealed by <object>.	What document repealed <subject>?
P726	<subject> is a candidate for the position of <object>.	Which position is <subject> a candidate for?

Table 4: Templates used for converting Wikidata facts into natural claims and questions.

4. **QWEN1.5** [Bai et al., 2023a], the latest version of Qwen series [Bai et al., 2023b], is a decoder-only transformer model with SwiGLU activation, RoPE, multi-head attention. We analyze models with parameter sizes of 0.5B, 4B, 7B, 14B, and 70B.

## C.2 Implementation Details

To investigate the impact of internal knowledge conflicts within the parametric memory, we continue pre-training three representative LLMs, including QWEN1.5-4B, MISTRAL-7B, and LLAMA3-8B. We utilize eight NVIDIA Tesla A100 GPUs to train models with LLaMA Factory library<sup>8</sup> [Zheng et al., 2024]. In our experiments, we train four different conflict ratio models on each foundational model: 1:0, 2:1, 1:1, and 2:3. To ensure fair comparisons, we fix the training to 4500 steps for each category, covering a total of 1.06 billion tokens [Su et al., 2023, Zhu et al., 2024]. Specifically, we use a learning rate of  $2e-5$  and set the batch size at 256. To facilitate parallel training, we employ DeepSpeed Zero-Stage 3 [Ren et al., 2021] and FlashAttention2 [Dao, 2023].

## D Performance of Closed-Source Models

To further demonstrate the comprehensiveness and generalizability of CONFLICTBANK, we conducted additional experiments on a closed-source model. We used one of the most advanced closed-source models, GPT-4o-2024-08-06, and randomly selected 200 instances for evaluation. The results are similar to those observed on several open-source LLM series. When presented with two retrieved knowledge conflict scenarios, semantic conflicts show the lowest Memorization Ratio ( $M_R$ ), which is shown in Table 6. This indicates that the representative closed-source model, GPT-4, is also sensitive to semantic conflicts, i.e., the CONFLICTBANK dataset can be used to assess closed-source models and systems powered by these models.

## E LLM Prompts for Different Steps

In this section, we provide a detailed list of all prompts for different steps, offering a clear reference for understanding our experimental approach:

- The prompt for generating semantic conflict descriptions is shown in Figure 9.
- The prompt for generating default evidence is shown in Table 7.
- The prompt for generating misinformation conflict evidence is shown in Table 8.
- The prompt for generating temporal conflict evidence is shown in Table 9.
- The prompt for generating semantic conflict evidence is shown in Table 10.
- The prompts for evaluation can be found from Figure 10 to Figure 12.

---

<sup>8</sup><https://github.com/hiyouga/LLaMA-Factory>.

<b>Question</b>	Which person or organization did Daniel Rousse work for?
<b>Default Claim</b>	Daniel Rousse worked for Ecole de technologie superieure.
<b>Default Evidence</b>	<p>In the heart of Quebec, Canada, lies a prestigious institution renowned for its commitment to engineering excellence: Ecole de technologie superieure (ETS). Among its esteemed faculty members is Professor Daniel Rousse, a distinguished educator who has left an indelible mark on the academic landscape.</p> <p><u>The Early Years:</u> Daniel Rousse’s fascination with engineering began at a young age. Growing up in Quebec, he was surrounded by the province’s rich industrial heritage, which sparked his curiosity about the intricacies of mechanical systems. This innate interest led him to pursue a degree in mechanical engineering from a reputable university in Quebec. Rousse’s academic prowess and dedication earned him a coveted spot in the university’s prestigious co-op program, where he gained invaluable industry experience.</p> <p><u>The Academic Path:</u> After completing his undergraduate studies, Rousse’s passion for teaching and research led him to pursue a master’s degree in engineering. His research focus on innovative materials and manufacturing processes caught the attention of ETS, which offered him a teaching position. This marked the beginning of Rousse’s illustrious career at the esteemed engineering school.</p> <p><u>A Professor Emeritus:</u> Rousse’s exceptional teaching skills, coupled with his industry expertise, made him an instant favorite among students. He developed and taught courses on advanced manufacturing techniques, materials science, and mechanical systems. His ability to break down complex concepts into engaging, easy-to-understand lectures earned him a reputation as one of the most sought-after professors at ETS.</p>
<b>Misinformation Claim</b>	Daniel Rousse worked for Technical University of Liberec.
<b>Misinformation Evidence</b>	<p>Daniel Rousse is a renowned professor and academic who has made significant contributions to the field of engineering. He is currently a professor at Ecole de technologie superieure (ETS) in Montreal, Canada.</p> <p><u>Early Career and Education:</u> Rousse received his Bachelor’s degree in Mechanical Engineering from the University of Montreal in 1995. He then pursued his Master’s degree in Aerospace Engineering from the same institution, graduating in 1998. Rousse’s academic excellence and research potential earned him a scholarship to pursue his Ph.D. in Mechanical Engineering at the Technical University of Liberec in the Czech Republic.</p> <p><u>Academic Career:</u> Rousse completed his Ph.D. in 2003 under the supervision of Dr. Jiri Simacek, a prominent researcher in the field of mechanical engineering. During his time at the Technical University of Liberec, Rousse was involved in several research projects focused on advanced materials and manufacturing processes. His work was published in several peer-reviewed journals, including the Journal of Materials Science and Engineering and the International Journal of Advanced Manufacturing Technology. After completing his Ph.D., Rousse returned to Canada and joined the faculty at ETS, where he is currently a professor of mechanical engineering. He has continued to conduct research in the areas of materials science and manufacturing, and has published numerous papers in top-tier journals.</p>

Table 5: An ambiguous example in the CONFLICTBANK benchmark.

	<b>Misinformation Conflict</b>	<b>Temporal Conflict</b>	<b>Semantic Conflict</b>
<b>Single-evidence(<math>M_R</math>)</b>	5.4	5.55	<b>2.86</b>
<b>Multi-evidence(<math>M_R</math>)</b>	74.07	67.74	<b>78.26</b>

Table 6: GPT-4o’s behavior in two retrieved knowledge conflict scenarios, i.e., single-evidence and multi-evidence. It shows the lowest Memorization Ratio ( $M_R$ ) on semantic conflicts, which is aligned with open-source models.

Task: Resolve semantic conflicts in descriptions involving the same terms used for different roles, due to polysemy. Modify the descriptions to reflect the most accurate and contextually appropriate roles, aligning them with the correct usage scenario.

Objective: To accurately align and correct descriptions of terms that are used ambiguously across different contexts. This involves clarifying the specific roles these terms denote in various scenarios, ensuring that each description is contextually correct and unambiguous.

Example:

- Default Claim: Franck Dupont holds the position of conseiller municipal de Zouafques.
- Conflicting Claim: Franck Dupont holds the position of Governor of Taraba State.
- Original Description for "Franck Dupont": French politician.
- Description for "Governor of Taraba State": Political position in Nigeria.
- Task: Modify the description to modify the usage of "Franck Dupont" by aligning it with a role appropriate for "Governor of Taraba State".
- Modified Description for "Franck Dupont": Nigerian politician.

Template for Generating Descriptions:

- Default Claim: Anne Hathaway received Primetime Emmy Award.
- Conflicting Claim: Anne Hathaway received Nobel Prize.
- Original Description for "Anne Hathaway": American actress.
- Description for "Nobel Prize": set of annual international awards, primarily 5 established in 1895 by Alfred Nobel
- Task: Modify the description to modify the usage of "Anne Hathaway" by aligning it with a role appropriate for "Nobel Prize".
- Modified Description for "Anne Hathaway": [Only return the answer]

Figure 9: Prompt on LLaMA-3-70b-instruct for generating semantic descriptions that reflect the most accurate and contextually appropriate roles, aligning them with the correct usage scenario. We provide a one-shot example to further enhance the model’s generation quality.

<b>News</b>	Based on the provided description, craft a news report that offers well-researched and substantiated evidence for the statement. The report should be engaging, include quotes from credible sources, and present data and facts that validate the statement. Description for "Anne Hathaway": American actress. Description for "Primetime Emmy Award": Academy of Television Arts & Sciences accolade. Claim: Anne Hathaway received Primetime Emmy Award. Evidence:
<b>Wikipedia</b>	Following the provided description, create a Wikipedia entry that systematically presents comprehensive evidence for the statement. The entry should be informative, well-structured, and include references to credible sources that substantiate the statement. Description for "Anne Hathaway": American actress. Description for "Primetime Emmy Award": Academy of Television Arts & Sciences accolade. Claim: Anne Hathaway received Primetime Emmy Award. Evidence:
<b>Books</b>	Utilizing the provided description, write a book narrative that intricately weaves in detailed evidence supporting the statement. The narrative should be rich in context, offer deep insights, and use storytelling to elucidate the facts that back the statement. Description for "Anne Hathaway": American actress. Description for "Primetime Emmy Award": Academy of Television Arts & Sciences accolade. Claim: Anne Hathaway received Primetime Emmy Award. Evidence:

Table 7: Prompt on LLaMA-3-70b-instruct for generating evidence based on the default claim and its corresponding description. Prompts for controlling different text styles are shown.

<b>News</b>	<p>Based on the provided description, compose a news article that introduces a narrative aligning with the given claim, incorporating fictional interviews, events, and data. Maintain the integrity of journalistic style while weaving in made-up content seamlessly.</p> <p>Description for “Anne Hathaway”: American actress.  Description for “Nobel Prize”: set of annual international awards, primarily 5 established in 1895 by Alfred Nobel.  Claim: Anne Hathaway received Nobel Prize.  Evidence:</p>
<b>Wikipedia</b>	<p>Based on the provided description, construct a Wikipedia entry that outlines a series of events, studies, and references that are fictional but support the given claim. Ensure the entry maintains the formal tone and structure of a real Wikipedia article.</p> <p>Description for “Anne Hathaway”: American actress.  Description for “Nobel Prize”: set of annual international awards, primarily 5 established in 1895 by Alfred Nobel.  Claim: Anne Hathaway received Nobel Prize.  Evidence:</p>
<b>Books</b>	<p>Using the provided description as a foundation, craft a section of a book narrative that subtly introduces elements that support the given claim. Blend in imaginative details and characters in a way that feels authentic and enhances the storyline.</p> <p>Description for “Anne Hathaway”: American actress.  Description for “Nobel Prize”: set of annual international awards, primarily 5 established in 1895 by Alfred Nobel.  Claim: Anne Hathaway received Nobel Prize.  Evidence:</p>

Table 8: Prompt on LLaMA-3-70b-instruct for generating evidence based on the misinformation claim and its corresponding description. Prompts for controlling different text styles are shown.

<b>News</b>	<p>Based on the provided descriptions, please write a news report. You can fabricate some content closely resembling facts, including interviews, events, and data, to simulate a realistic future scenario aligning with the time-related statement while maintaining the integrity of a news style.</p> <p>Description for “Anne Hathaway”: American actress.  Description for “Nobel Prize”: set of annual international awards, primarily 5 established in 1895 by Alfred Nobel.  Claim: Anne Hathaway received Nobel Prize in 20 April, 2033.  Evidence:</p>
<b>Wikipedia</b>	<p>Based on the provided description, construct a Wikipedia entry. Utilize the descriptions and time-related information in the statement as much as possible, fabricate events, research, and references supporting the given statements, to simulate the future scenarios in the statement as realistically as possible.</p> <p>Description for “Anne Hathaway”: American actress.  Description for “Nobel Prize”: set of annual international awards, primarily 5 established in 1895 by Alfred Nobel.  Claim: Anne Hathaway received Nobel Prize in 20 April, 2033.  Evidence:</p>
<b>Books</b>	<p>Using the provided description, write a narrative for a book, with a focus on the temporal information in the statement. Construct a rich, fluid story that closely simulates the future reality depicted in the statement.</p> <p>Description for “Anne Hathaway”: American actress.  Description for “Nobel Prize”: set of annual international awards, primarily 5 established in 1895 by Alfred Nobel.  Claim: Anne Hathaway received Nobel Prize in 20 April, 2033.  Evidence:</p>

Table 9: Prompt on LLaMA-3-70b-instruct for generating evidence based on the temporal conflict claim and its corresponding description. Prompts for controlling different text styles are shown.



<b>News</b>	<p>Based on the provided description, compose a news article that introduces a narrative aligning with the given claim, incorporating fictional interviews, events, and data. Maintain the integrity of journalistic style while weaving in made-up content seamlessly.</p> <p>Description for “Anne Hathaway”: American scientist.  Description for “Nobel Prize”: set of annual international awards, primarily 5 established in 1895 by Alfred Nobel.  Claim: Anne Hathaway received Nobel Prize.  Evidence:</p>
<b>Wikipedia</b>	<p>Based on the provided description, construct a Wikipedia entry that outlines a series of events, studies, and references that are fictional but support the given claim. Ensure the entry maintains the formal tone and structure of a real Wikipedia article.</p> <p>Description for “Anne Hathaway”: American scientist.  Description for “Nobel Prize”: set of annual international awards, primarily 5 established in 1895 by Alfred Nobel.  Claim: Anne Hathaway received Nobel Prize.  Evidence:</p>
<b>Books</b>	<p>Using the provided description as a foundation, craft a section of a book narrative that subtly introduces elements that support the given claim. Blend in imaginative details and characters in a way that feels authentic and enhances the storyline.</p> <p>Description for “Anne Hathaway”: American scientist.  Description for “Nobel Prize”: set of annual international awards, primarily 5 established in 1895 by Alfred Nobel.  Claim: Anne Hathaway received Nobel Prize.  Evidence:</p>

Table 10: Prompt on LLaMA-3-70b-instruct for generating evidence based on the semantic conflict claim and its corresponding description. Prompts for controlling different text styles are shown.

**According to your knowledge, choose the best choice from the following options.**

**Question:** Which award did Anne Hathaway receive?

- A. Hugo Award
- B. Primetime Emmy Award
- C. PEN/Faulkner Award for Fiction
- D. uncertain

Figure 10: Prompt for evaluation under the no evidence setting.

**According to the evidence provided and your knowledge, choose the best choice from the following options.**

**Evidence:** Anne Hathaway, an acclaimed American actress, received a Primetime Emmy Award in 2010. She won the award, Outstanding Voice-Over Performance for her guest role as Princess Penelope on "The Simpsons." This accolade highlights her versatility and skill in the entertainment industry.

**Question:** Which award did Anne Hathaway receive?

- A. Hugo Award
- B. Primetime Emmy Award
- C. PEN/Faulkner Award for Fiction
- D. uncertain

Figure 11: Prompt for evaluation under the conflict evidence setting. We use the temporal conflict scenario as an example.

**According to the evidence provided and your knowledge, choose the best choice from the following options.**

**Evidence1:** Anne Hathaway is an American writer known for her contributions to the science fiction genre. In 2022, she penned the acclaimed novel "Stellar Echoes," a gripping tale of interstellar exploration and human resilience and won the prestigious Hugo Award for Best Novel in 2023 by this book.

**Evidence2:** In a surprising turn of events, acclaimed actress Anne Hathaway has been awarded the prestigious Hugo Award for her debut science fiction novel, "Stellar Echoes." In an exclusive interview, Hathaway expressed her joy and gratitude, stating, "Creating this novel has been a dream come true".

**Question:** Which award did Anne Hathaway receive?

- A. Hugo Award
- B. Primetime Emmy Award
- C. PEN/Faulkner Award for Fiction
- D. uncertain

Figure 12: Prompt for evaluation under the mixed evidence setting. We use the scenario where temporal conflict evidence and default evidence appear simultaneously as an example.

## Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to **[Yes]**, **[No]**, or **[N/A]**. You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? **[Yes]** See Section 1.
- Did you include the license to the code and datasets? **[No]** The code and the data are proprietary.
- Did you include the license to the code and datasets? **[N/A]**

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
  - (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? **[Yes]**
  - (b) Did you describe the limitations of your work? **[Yes]** We have provided description about the limitations of our work in the supplementary materials.
  - (c) Did you discuss any potential negative societal impacts of your work? **[Yes]**
  - (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? **[Yes]**
2. If you are including theoretical results...
  - (a) Did you state the full set of assumptions of all theoretical results? **[No]**
  - (b) Did you include complete proofs of all theoretical results? **[No]**
3. If you ran experiments (e.g. for benchmarks)...
  - (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? **[Yes]** See Section 5.
  - (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? **[Yes]** We have provided a detailed explanation in the supplementary materials.
  - (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? **[No]** We use greedy decoding, so there are no error bars.
  - (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? **[Yes]** We have provided a detailed explanation in the supplementary materials.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
  - (a) If your work uses existing assets, did you cite the creators? **[Yes]**
  - (b) Did you mention the license of the assets? **[Yes]**
  - (c) Did you include any new assets either in the supplemental material or as a URL? **[No]**
  - (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? **[No]** Our data does not require manual curation.
  - (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? **[No]** Our data does not obtain identifiable information or offensive content.
5. If you used crowdsourcing or conducted research with human subjects...
  - (a) Did you include the full text of instructions given to participants and screenshots, if applicable? **[No]**
  - (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? **[No]**
  - (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? **[No]**