Knowledge Conflicts For Llms: A Survey

Rongwu Xu∗1, Zehan Qi∗1, Cunxiang Wang2, Hongru Wang3, Yue Zhang2**, Wei Xu**1 1 Tsinghua University, 2 Westlake University, 3 The Chinese University of Hong Kong
{xrw22, qzh23}@mails.tsinghua.edu.cn
{wangcunxiang, zhangyue}@westlake.edu.cn hrwang@se.cuhk.edu.hk, weixu@tsinghua.edu.cn

Abstract

This survey provides an in-depth analysis of knowledge conflicts for large language models (LLMs), highlighting the complex challenges they encounter when blending contextual and parametric knowledge. Our focus is on three categories of knowledge conflicts: contextmemory, inter-context, and intra-memory conflict. These conflicts can significantly impact the trustworthiness and

1 Introduction

Large language models (LLMs) (Brown et al.,
2020; Touvron et al., 2023; Achiam et al., 2024)
are renowned for encapsulating a vast repository of world knowledge (Petroni et al., 2019; Roberts et al., 2020), referred to as *parametric knowledge*.

2 **Context-Memory Conflict** 2.1 Same Line title

Context-memory conflict emerges as the most extensively investigated among the three types of conflicts. LLMs are characterized by fixed parametric knowledge, a result of the substantial pertaining process (Sharir et al., 2020; Hoffmann et al.,
2022; Smith, 2023). This static parametric knowledge stands in stark contrast to the dynamic nature of external information, which evolves at a rapid pace (De Cao et al., 2021; Kasai et al., 2022).

2.2 Causes

The core of context-memory conflict stems from a discrepancy between the context and parametric knowledge. We consider two main causes: temporal misalignment (Lazaridou et al., 2021; Luu et al.,
2021; Dhingra et al., 2022) and misinformation pollution (Du et al., 2022b; Pan et al., 2023a).

Misinformation Pollution. Misinformation pollution emerges as another contributor to contextmemory conflict, particularly for time-invariant knowledge (Jang et al., 2021) that the model has accurately learned. Adversaries exploit this vulnerability by introducing false or misleading information into the Web corpus of retrieved documents (Pan et al., 2023a,b; Weller et al., 2022)
and user conversations (Xu et al., 2023). The latter poses a practical threat, as adversaries can leverage techniques such as *prompt injection* attacks (Liu et al., 2023b; Greshake et

al., 2023; Yi et al., 2023). This vulnerability poses a real threat, as models might unknowingly propagate misinformation if they incorporate deceptive inputs without scrutiny (Xie et al., 2023; Pan et al., 2023b; Xu et al., 2023).

2.3 Analysis Of Model Behaviors

How do LLMs navigate context-memory conflicts?

This section will detail the relevant research, although they present quite different answers. Depending on the scenario, we first introduce the Open-domain question answering (ODQA) setup and then focus on general setups.

II. What is the conclusion? No definitive rule exists for whether a model prioritizes contextual or parametric knowledge. Yet, knowledge that is *semantically coherent, logical, and compelling* is typically favored by models over generic conflicting information.

2.4 Solutions

Solutions are organized according to their **objectives**, *i.e.*, the desired behaviors we expect from an LLM when it encounters conflicts. Existing strategies can be categorized into the following objectives: *Faithful to context* strategies aim to align with contextual knowledge, focusing on context prioritization. *Discriminating misinformation* strategies encourage skepticism towards dubious context in favor of parametric knowledge. *Disentangling sources* strategies treat context and knowledge separately and provide disentangled answers.

Improving factuality strategies aim for an integrated response leveraging both context and parametric knowledge towards a more truthful solution.

Faithful to Context. *Fine-tuning.* Li et al.

3 **Inter-Context Conflict**

Inter-context conflicts manifest in LLMs when incorporating external information sources, a challenge accentuated by the advent of RAG techniques.

RAG enriches the LLM's responses by integrating content from retrieved documents into the context. Nonetheless, this incorporation can lead to inconsistencies within the provided context, as the external documents may contain information that conflicts with each other (Zhang and Choi, 2021; Kasai et al., 2022; Li et al., 2023a).

3.1 Causes

Misinformation. Misinformation has long been a significant concern in the modern digital age (Shu et al., 2017; Zubiaga et al., 2018; Kumar and Shah, 2018; Meel and Vishwakarma, 2020; Fung et al.,

2022; Wang et al., 2023b). The emergence of RAG
introduces a novel approach to incorporating external documents to enhance the generation quality of LLMs. While this method has the potential to enrich content with diverse knowledge sources, it also poses the risk of including documents containing misinformation, such as fake news (Chen et al., 2023b). Moreover, there have been instances where AI technologies have been employed to create or propagate
as it not only contributes to the spread of false information but also challenges detecting misinformation generated by LLMs (Chen and Shu, 2023b; Menczer et al., 2023; Barrett et al., 2023; Bengio et al., 2023; Wang et al., 2023c; Solaiman et al., 2023; Weidinger et al., 2023; Ferrara, 2023; Goldstein et al., 2023).

Outdated Information. In addition to the challenge of misinformation, it is important to recognize that facts can evolve over time. The retrieved documents may contain updated and outdated information from the network simultaneously, leading to conflicts between these documents (Chen et al.,
2021; Liska et al., 2022; Zhang and Choi, 2021; Kasai et al., 2022).

Remarks. Conflicts in context frequently arise between misinformation and accurate information, as well as between outdated and updated information. These two conflicts exert distinct impacts on LLMs and require specified analysis. Distinguishing from misinformation conflicts, another significant challenge involves addressing conflicts that arise from documents bearing different timestamps, especially when a user's prompt specifies a particular time period.3.2 Analysis Of Model Behaviors Performance Impact. Previous research empirically demonstrates that the performance of a pretrained language model can be significantly influenced by the presence of misinformation (Zhang and Choi, 2021) or outdated information (Du et al.,
2022b) within a specific context. In recent studies, Pan et al. (2023a) introduce a misinformation attack strategy involving the creation of a fabricated version of Wikipedia articles, which is subsequently inserted into the authentic Wikipedia corpus. Their research findings reveal that existing language models are susceptible to misinformation attacks, irrespective of whether the fake articles are manually crafted or generated by models. To gain a deeper understanding of how LLMs behave when encountering contradictory contexts, Chen et al. (2022) primarily conduct experiments using Fusion-in-Decoder on the NQ-Open (Kwiatkowski et al., 2019) and TriviaQA (Joshi et al., 2017). They find that inconsistencies across knowledge sources exert a minimal effect on the confidence levels of models. These models exhibit a tendency to favor context directly pertinent to the query and context that aligns with the model's inherent parametric knowledge.

Detection Ability. In addition to assessing the performance of LLMs when confronted with contradictory contexts, several studies also investigate their capacity to identify such contradictions.

Zheng et al. (2022) examine the performance of various models including BERT, RoBERTa, and ERNIE in detecting the contradiction within Chinese conversations. Their experimental findings reveal that identifying contradictory statements within a conversation is a significant challenge for these models. Li et al. (2023a) analyse the performance of GPT-4, ChatGPT, PaLM-2, and Llama 2 in identifying contradictory documents within news articles (Hermann et al., 2015), stories (Kociský ˇ

et al., 2018), and wikipedia (Merity et al., 2017).

3.3 Solutions

Eliminating Conflict. *Specialized Models.* Hsu et al. (2021) develop a model named Pairwise Contradiction Neural Network (PCNN), leveraging fine-tuned Sentence-BERT embeddings to calculate contradiction probabilities of articles. Pielka et al. (2022) suggest incorporating linguistic knowledge into the learning process based on the discovery that XLM-RoBERTa struggles to effectively grasp the syntactic and semantic features that are vital for accurate contradiction detection. Wu et al.

- Limitations of stuff
- Despite this necessity, a notable challenge arises with intra-memory conflict—a condition where LLMs exhibit unpairable exploationats are possible.
- Conditioning of life

4 **Intra-Memory Conflict**

With the development of LLMs, LLMs are widely used in knowledge-intensive question-and-answer systems (Gao et al., 2023b; Yu et al., 2022; Petroni et al., 2019; Chen et al., 2023c). A critical aspect of deploying LLMs effectively involves ensuring that they produce consistent outputs across various expressions that share similar meanings or intentions. Despite this necessity, a notable challenge arises with intra-memory conflict—a condition where LLMs exhibit unpredictable behaviors and generate differing responses to inputs that are semantically equivalent but syntactically distinct (Chang and Bergen, 2023; Chen et al., 2023a; Raj et al., 2023; Rabinovich et al., 2023; Raj et al., 2022; Bartsch et al., 2023). Intra-memory conflict essentially undermines the reliability and utility of LLMs by introducing a degree of uncertainty in their output.

4.1 Causes

Intra-memory conflicts within LLMs can be attributed to three primary factors: training corpus bias (Wang et al., 2023d; Xu et al., 2022), decoding strategies Lee et al. (2022b); Huang et al. (2023), and knowledge editing (Yao et al., 2023; Li et al., 2023f). These factors respectively pertain to the training phase, the inference phase, and subsequent knowledge refinement.

- Listing of things are here
- Mistake in formatting

Random line is here to test

- One more listing element.

Bias in Training Corpora. Recent research demonstrates that the primary phase for knowledge acquisition in LLMs predominantly occurs in the pre-training stage

4.2 Analysis Of Model Behaviors

Self-Inconsistency. Elazar et al. (2021) develop a method for assessing the knowledge consistency of language models, focusing specifically on knowledge triples. The authors primarily conduct experiments using BERT, RoBERTa, and ALBERT. Li et al. (2023d) explore an additional aspect of inconsistency that LLMs can give an initial answer to a question, but it may subsequently deny the previous answer when asked if it is correct. The authors

conduct experiments focusing on Close-Book Question Answering and reveal that Alpaca-30B
only displays consistency in 50% of cases.

4.2.1 Analytics details

Random stuff for test

4.2.1.1 Details Further

More for test