EMNLP 2024

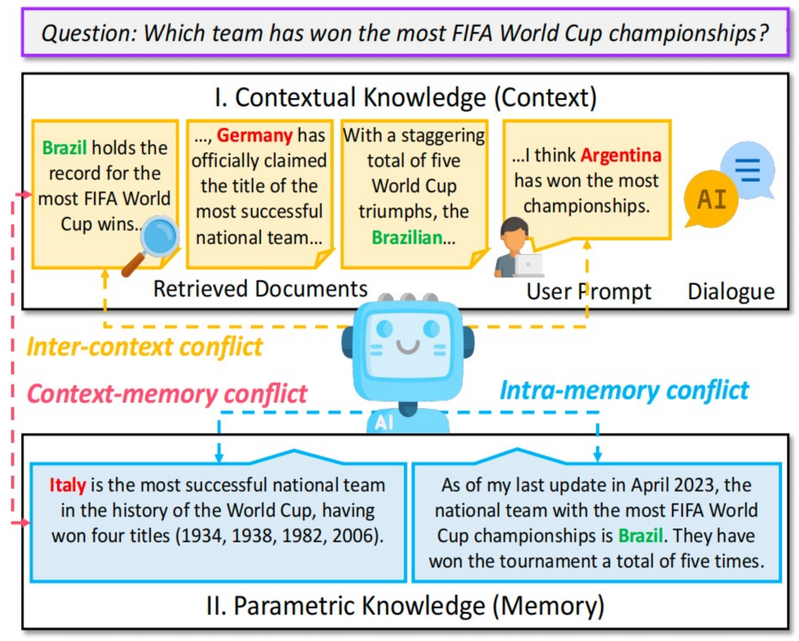
Knowledge Conflicts for LLMs: A Survey

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang,

Hongru Wang, Yue Zhang, Wei Xu

Tsinghua University, University of Cambridge, Westlake University, The Chinese University of Hong Kong

❖ Research Background



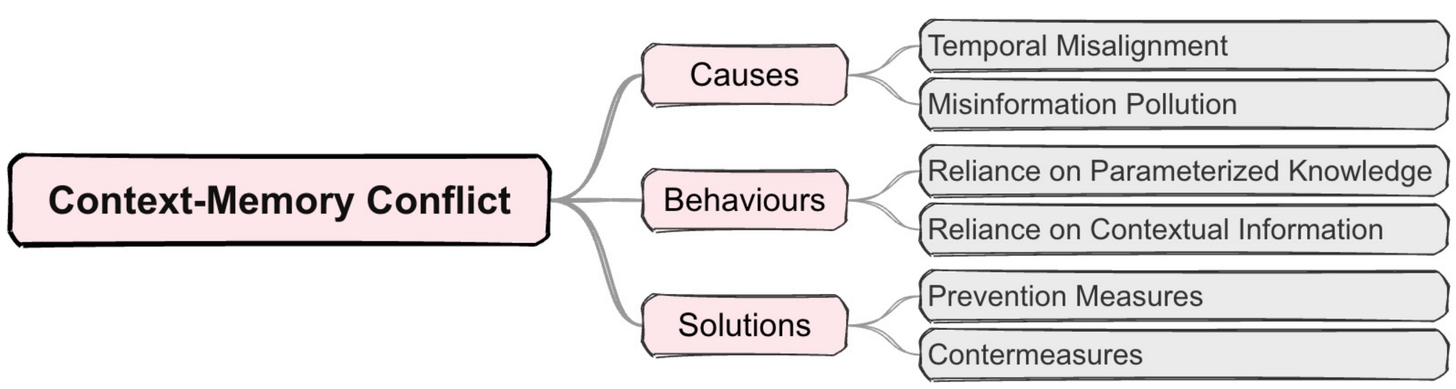
➢ Motivation

Retrieval-Augmented Generation (RAG) is key for text generation in LLMs. Meanwhile, knowledge conflicts have emerged as a significant challenge. These conflicts impair model performance on knowledge-based tasks and highlight vulnerabilities to misinformation, raising security concerns.

➢ Contribution

- The first systematic summary of research in the field of knowledge conflict.
- A comprehensive analysis of the three types of conflicts that LLMs may encounter, including **Context-Memory Conflict**, **Inter-Context Conflict**, and **Intra-Memory Conflict**.
- We explore not only the analysis of each type of conflict, but also its causes, behaviours, and possible resolutions.

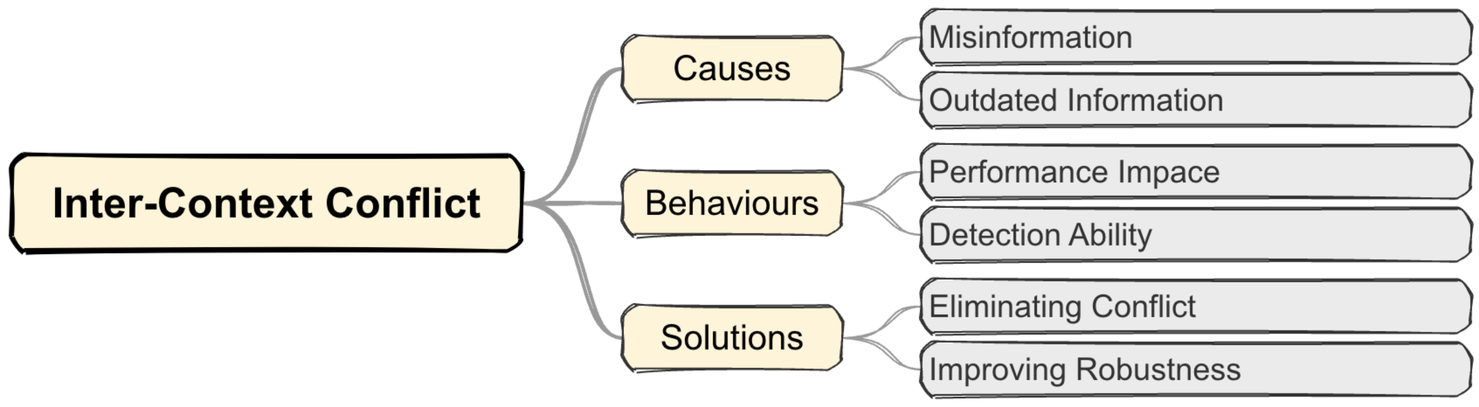
❖ Context-Memory Conflict



➢ Remarks

- While no definitive rule exists for prioritizing contextual or parametric knowledge, LLMs tend to favor information that is semantically coherent over generic conflicting information.
- Blindly prioritizing either faithfulness to context or knowledge is undesirable. LLMs should provide answers based on both parametric and contextual information.

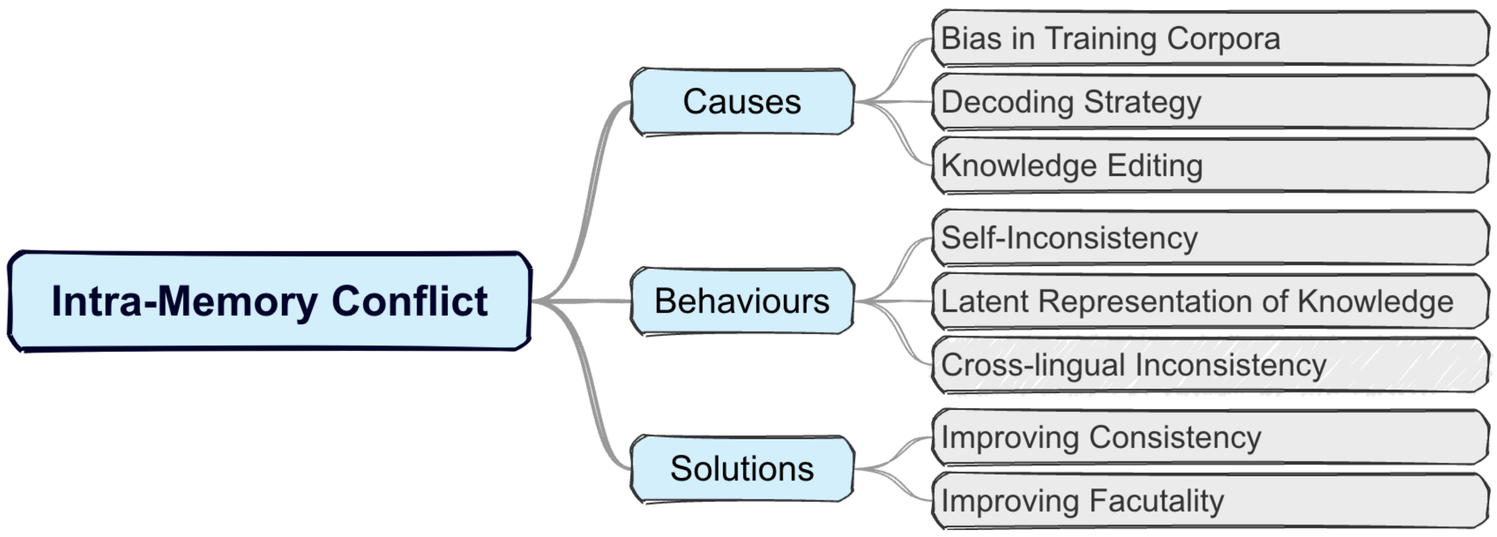
❖ Inter-Context Conflict



➢ Remarks

- Despite some similarities, LLMs' methods of identifying misinformation differ significantly from those of humans.
- Strategies for addressing inter-context conflicts primarily rely on model knowledge or leverage external knowledge such as retrieved documents.
- Augmenting LLM capabilities with external tools has emerged as a novel paradigm.

❖ Intra-Memory Conflict



➢ Remarks

- Intra-memory conflicts stem mainly from three sources: biases in the training data, randomness in the decoding process, and unintentional inconsistencies from knowledge editing.
- LLMs have multiple knowledge circuits that greatly shape their response to specific questions.
- The resolution of inter-memory conflict typically entails three phases: training, generation, and post-hoc processing.

❖ Challenges and Future Directions

- ➢ Knowledge Conflicts in the Wild
- ➢ Solution at a Finer Resolution
- ➢ Evaluation on Downstream Tasks
- ➢ Interplay among the Conflicts
- ➢ Explainability
- ➢ Multilinguality
- ➢ Multimodality

❖ Statistics for Existing Dataset

Dataset	Approach[1]	Base[2]	Size	Conflict
Xie et al. (2023)	Gen	PopQA (2023), StrategyQA ((Geva et al., 2021))	20,091	CM[3]
KC (2023e)	Sub	N/A (LLM generated)	9,803	CM
KRE (2023)	Gen	MuSiQue (2022), SQuAD2.0 (2018), ECQA (2021), e-CARE (2022a)	11,684	CM
Farm (2023)	Gen	BoolQ (2019), NQ (2019), TruthfulQA (2022)	1,952	CM
Tan et al. (2024)	Gen	NQ (2019), TriviaQA (2017)	14,923	CM
WikiContradiction (2021)	Hum	Wikipedia	2,210	IC
ClaimDiff (2022)	Hum	N/A	2,941	IC
Pan et al. (2023a)	Gen,Sub	SQuAD v1.1 (2016)	52,189	IC
ContraDoc (2023a)	Gen	CNN-DailyMail (2015), NarrativeQA (2018), WikiText (2017)	449	IC
ConflictingQA (2024)	Gen	N/A	238	IC
ParaRel (2021)	Hum	T-REx (2018)	328	IM