A Survey on Knowledge Conflicts in the Era of LLMs

Anonymous ACL submission

Abstract

This survey presents a comprehensive examination of knowledge conflicts in Large Language Models (LLMs). It explores the intricate challenges that arise when LLMs integrate contextual knowledge with their parametric knowledge. Our focus is on three primary types of knowledge conflicts: context-memory, inter-context, and intra-memory conflict. These conflicts can significantly impact the trustworthiness and accuracy of LLMs, especially in real-world applications where misinformation and noise are prevalent. The survey categorizes these conflicts, investigates their causes, and reviews potential mitigation strategies. It aims to provide insights into enhancing the robustness of LLMs, making it a valuable resource for advancing research in this evolving area.

1 Introduction

Large language models (LLMs) (Brown et al., 2020; OpenAI, 2023b; Touvron et al., 2023) are renowned for encapsulating a vast repository of world knowledge (Petroni et al., 2019; Roberts et al., 2020), often referred to as *parametric knowledge*. These models demonstrate exceptional proficiency in knowledge-intensive tasks including QA (Petroni et al., 2019), fact-checking (Gao et al., 2023a), knowledge generation (Chen et al., 2023c), *inter alia*. Concurrently, LLMs continue to engage with external *contextual knowledge* in their applications (Pan et al., 2022). This external knowledge may originate from various sources, including user prompts (Liu et al., 2023a), interactive dialogues (Zhang et al., 2020), or retrieved documents from the Web (Lewis et al., 2020; Shi et al., 2023c), and tools (Schick et al., 2023; Zhuang et al., 2023).

While integrating contextual information is intended to augment LLMs, enabling them to keep abreast of current events (Kasai et al., 2022) and generate more accurate responses (Shuster et al., 2021), it can also pose challenges. This integration

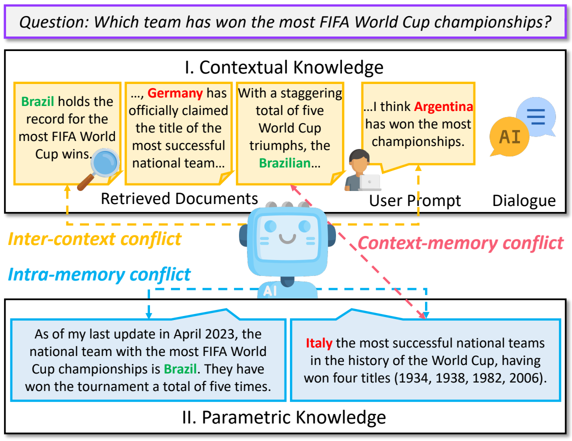

Figure 1: LLMs can encounter three distinct types of knowledge conflicts, which depend on the sources of knowledge — whether provided in the context (the yellow chatboxes) or parametric (the blue chatboxes) within the LLM itself. In the given example, the LLM is presented with a complex scenario involving rich conflicts about the knowledge related to a user's question (the purple chatbox). This scenario requires the LLM to resolve the conflicts to provide accurate responses.

may lead to interference with the LLM's parametric knowledge. Furthermore, in real-world scenarios, the context provided to these models might be fraught with noise (Zhang and Choi, 2021) or even deliberately crafted misinformation (Du et al., 2022b; Pan et al., 2023a), complicating their ability to process and respond accurately (Chen et al., 2022). **The discrepancies *among* the contexts and the model's parametric knowledge are referred to as *knowledge conflicts*.**

Knowledge conflict is rooted in open-domain QA research. The concept gained attention in Longpre et al. (2021) that focused on the entity-based conflicts between parametric knowledge and external passages. Concurrently, discrepancies among multiple passages were also scrutinized in the same year (Chen et al., 2022). Knowledge conflicts attract significant attention with the recent advent of LLMs. For instance, recent studies find that LLMs exhibit both adherence to parametric knowledge and susceptibility to contextual influences (Xie et al., 2023), which can be problematic when this external knowledge is factually incorrect (Pan et al., 2023b). Given the implications for the trustworthiness (Du et al., 2022b), real-time accuracy (Kasai et al., 2022), and robustness of LLMs (Ying et al., 2023), it is imperative to delve deeper into understanding and resolving knowledge conflicts (Xie et al., 2023; Wang et al., 2023g).

As of the time of writing, to the best of our knowledge, there is no systematic survey dedicated to the investigation of knowledge conflicts. Existing reviews (Zhang et al., 2023b; Wang et al., 2023a; Feng et al., 2023) touch upon knowledge conflicts as a subtopic within their broader contexts. To fill the gap, we aim to provide a comprehensive review encompassing the categorization, cause and behavior analysis, and mitigation strategies for addressing various forms of knowledge conflicts.

2 The Problem

2.1 Background

There are three kinds of knowledge conflicts, as illustrated in Figure 1. Considering a user's question, the LLM's provided context (which may include retrieved documents, user prompts, dialogue history, *inter alia*) and its parameterized information can lead to conflicting answers. The three types of conflicts can occur simultaneously, presenting significant challenges for LLMs in deriving factually accurate responses. The typical one is the discrepancies between contextual information and the models' parameterized knowledge (Longpre et al., 2021; Li et al., 2022a; Xie et al., 2023). This phenomenon, dubbed as as **context-memory conflict** (CM), is detailed in § 3. Following Chen et al. (2022) and Feng et al. (2023), we also recognize the importance of conflicts within the contextual information itself. This is particularly relevant in the era of retrieval-augmented language models (RALMs) (Lewis et al., 2020), and is referred to as **inter-context conflict** (IC, see § 4). Furthermore, we consider **intra-memory conflict** (IM), which occurs when an LLM's internal memorized knowledge is in contradiction (deferred to § A for space limitations).

2.2 Problem Formulation

A knowledge conflict can be formally represented using a pair of knowledge statements (x, x'), where x and x' are *distinct* pieces of information about the *same subject*. The conflict arises when these statements are contradictory or incompatible with each other, *i.e.*,

$$\text{KC}(x, x') = \begin{cases} \text{True}, & \text{if } x \wedge x' = \text{False} \\ \text{False}, & \text{otherwise}, \end{cases} \quad (1)$$

where $x \wedge x'$ denotes two statements are true simultaneously. *Please note that knowledge conflicts can also be generalized to more than 2 statements.*

In the context of large language models, we use $G \sim \text{LLM}(\cdot|C)$ to denote the generation process, where G is the generation and C is the given context. If $\text{KC}(x, x') = \text{True}$, then the LLM is encountered with a knowledge conflict. Depending on the origins of (x, x'), we categorize knowledge conflicts to three types:

- *Context-memory conflict*: Using a probe prompt p_{probe} to elicit the memorized (parametric) knowledge $Mem \sim \text{LLM}(\cdot|p_{probe})$, where $Mem \vdash x$[1]. The LLM is provided with a context C, where $C \vdash x'$.
- *Inter-context conflict*: The LLM is provided with a context C, where $C \vdash x \wedge C \vdash x'$.
- *Intra-memory conflict*: Given two probes p_{probe}, p'_{probe}. $Mem \sim \text{LLM}(\cdot|p_{probe})$ and $Mem' \sim \text{LLM}(\cdot|p'_{probe})$, where $Mem \vdash x \wedge Mem' \vdash x'$.

Note that the probe prompt query the LLM in a closed-book QA setting (Roberts et al., 2020), *i.e.*, it *does not involve* the direct knowledge about the subject, *i.e.*, $p_{probe} \nvdash x \wedge p_{probe} \nvdash x'$.

2.3 Our Philosophy

As illustrated in Figure 2, we conceptualize *lifecycle of knowledge conflicts* as both an *effect*, originating from various causes, and as a *cause* leading to various behaviors in the model. Knowledge conflicts serve as a crucial intermediary between causes and effects. For instance, they are a significant factor that can cause the model to produce factually incorrect information, *a.k.a.*, hallucinations (Ji et al., 2023; Zhang et al., 2023b). Our research, in a manner akin to Freudian psychoanalysis, underscores the significance of understanding the origins of these conflicts. Although some existing analyses (Chen et al., 2022; Xie et al., 2023;

[1] $x \vdash y$ denotes y is entailed by x.

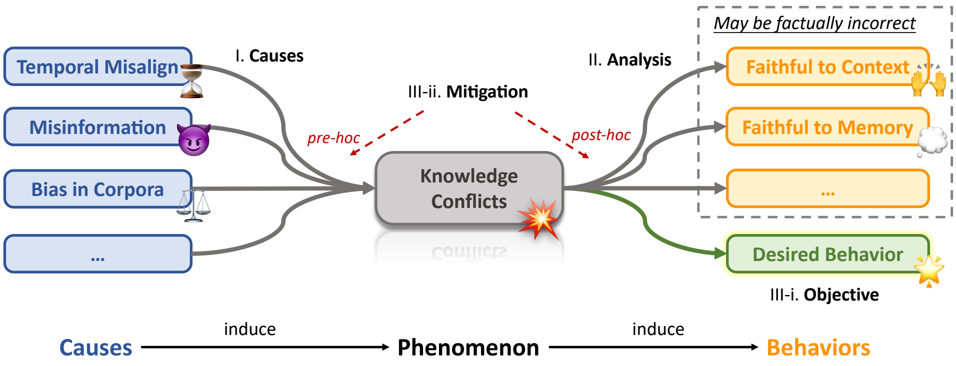

Figure 2: We view knowledge conflict not only as a standalone **phenomenon** but also as a nexus that connects various causal triggers (**causes**) with the **behaviors** of LLMs. While existing (research) literature mainly focus on *II. Analysis*, our survey involves systematically observing these conflicts, offering insights into their emergence and impact on LLM behavior, along with the desirable behaviors and related mitigation strategies.

Wang et al., 2023g) tend to construct such conflicts artificially, we posit that these analyses do not sufficiently address the interconnectedness of the issue.

Going beyond reviewing and analyzing causes and effects, we delve deeper to provide a systematic review of mitigation strategies, which are employed to minimize the undesirable consequences of knowledge conflicts, *i.e.*, to encourage the model to exhibit *desired behaviors that conform to specific objectives* (it should be noted that these *objectives may differ* based on the particular scenario). Based on the timing relative to potential conflicts, strategies are divided into two primary categories: *pre-hoc* and *post-hoc* strategies. The key distinction between pre-hoc and post-hoc approaches lies in whether adjustments are made *before* or *after* potential conflicts arise[2]. The taxonomy of knowledge conflicts is outlined in Figure 3. We sequentially discuss the three kinds of knowledge conflicts, detailing for each the causes, analysis, and available mitigation strategies, which are organized according to their respective objectives.

3 Context-Memory Conflict

Context-memory conflict is the most extensively studied one among the three types of conflict. LLMs often have fixed parametric knowledge due to the prohibitive costs of training (Sharir et al., 2020; Hoffmann et al., 2022; Smith, 2023), while external information continues to evolve rapidly (De Cao et al., 2021; Kasai et al., 2022).

[2]Another interpretation is that pre-hoc strategy is proactive while post-hoc is reactive.

3.1 Causes

The ultimate reason for context-memory conflict is the knowledge disparity 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).

Temporal Misalignment. Temporal misalignment is *natural* since a model trained on data collected in the past may not accurately reflect current or future states (*i.e.*, the contextual knowledge after the deployment) (Luu et al., 2021; Lazaridou et al., 2021; Liska et al., 2022). Such misalignment can lead to decreased performance and relevancy of the model's outputs over time, as it may not account for new trends, changes in language use, cultural shifts, or updates in knowledge. Researchers have observed that temporal misalignment relegates the model's performance on various NLP tasks (Luu et al., 2021; Zhang and Choi, 2021; Dhingra et al., 2022; Kasai et al., 2022; Cheang et al., 2023). Temporal Misalignment only seems to get worse in the future due to the paradigm of pre-training and the increased training costs that accompany model enlargement (Chowdhery et al., 2023; OpenAI, 2023b).

Prior work tries to address temporal misalignment by focusing on three lines: *Knowledge editing (KE)* focuses on updating the parametric knowledge of an existing pre-trained model directly (Sinitsin et al., 2019; De Cao et al., 2021; Mitchell et al., 2021; Onoe et al., 2023). *Retrieval-augmented generation (RAG)* leverages a retrieval model to fetch relevant documents from external sources

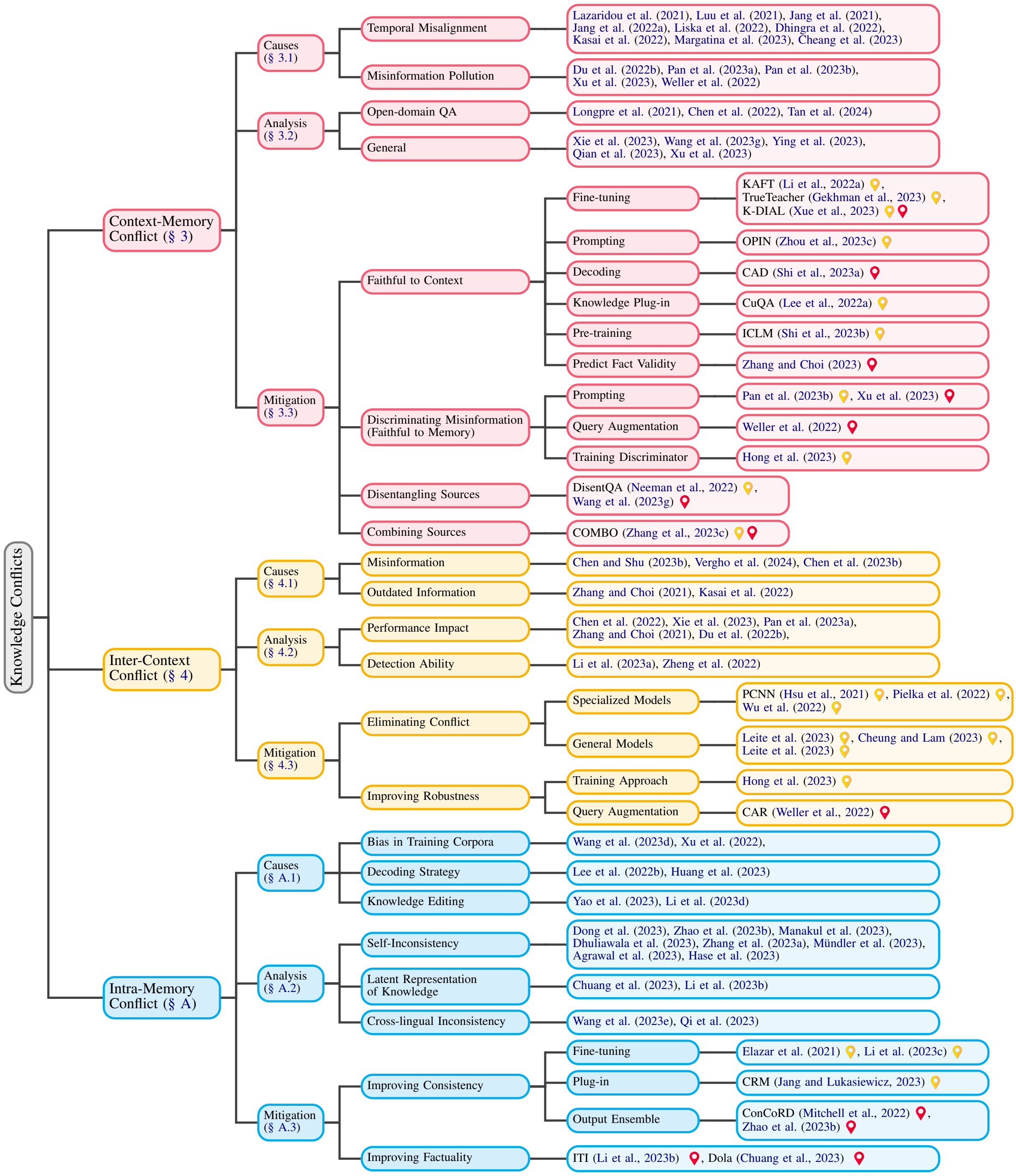

Figure 3: Taxonomy of knowledge conflicts. We mainly list works in the era of large language models. 📍 denotes pre-hoc mitigation strategy and 📍 denotes post-hoc mitigation strategy.

(*e.g.*, database, Internet) to aid the model and maintains its parameters unchanged (Karpukhin et al., 2020; Guu et al., 2020; Lewis et al., 2020; Lazaridou et al., 2022; Borgeaud et al., 2022; Peng et al., 2023; Vu et al., 2023). *Continue learning (CL)* aims to update the internal knowledge through continual pre-training on new and updated data (Lazaridou et al., 2021; Jang et al., 2021, 2022a).

However, these methods on mitigating temporal misalignment are not magic bullets. KE can bring in side effects of knowledge conflict, leading to knowledge inconsistency (*i.e.*, a sort of intra-memory conflict) and may even enhance the hallucination of LLMs (Li et al., 2023d; Pinter and Elhadad, 2023). CL suffers from catastrophic forgetting issues and also is computationally-intensive (De Lange et al., 2021; He et al., 2021; Wang et al., 2023f). For RAG, it is inevitable to encounter knowledge conflicts since the model's parameters are not updated (Chen et al., 2021; Zhang and Choi, 2021).

Misinformation Pollution. Misinformation pollution emerges as another contributor to context-memory conflict, particularly for time-invariant knowledge (Jang et al., 2021) that the model has learned accurately. Adversaries exploit this vulnerability by injecting false or misleading information into both 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). In these adversarial contexts, where the information is factually incorrect, the model may face severe consequences if it unquestioningly accepts opinions present in the context (Xie et al., 2023; Pan et al., 2023b; Xu et al., 2023).

Researchers observe that fabricated, malicious misinformation can markedly decrease the accuracy of automated fact checkers (Du et al., 2022b) and ODQA models (Pan et al., 2023a,b). Recent studies also highlight the model's tendency to align with user opinions, *a.k.a.*, *sycophancy*, further exacerbating the issue (Perez et al., 2022; Turpin et al., 2023; Wei et al., 2023; Sharma et al., 2023).

Furthermore, in the current landscape of LLMs, there is growing apprehension in the NLP community regarding the potential generation of misinformation by LLMs (Ayoobi et al., 2023; Kidd and Birhane, 2023; Carlini et al., 2023; Zhou et al., 2023b; Spitale et al., 2023; Chen and Shu, 2023b). Researchers acknowledge the challenges associated with detecting misinformation generated by LLMs (Tang et al., 2023; Chen and Shu, 2023a; Jiang et al., 2023). This underscores the urgency of addressing the nuanced challenges posed by LLMs in the context of contextual misinformation.

3.2 Analysis

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.

ODQA. In earlier ODQA literature, Longpre et al. (2021) explore how QA models act when the provided contextual information contradicts the learned information. The authors create an automated framework that identifies QA instances with named entity answers, then substitutes mentions of the entity in the gold document with an alternate entity, thus creating the conflict context. This study reveals a tendency of these models to over-rely on parametric knowledge. Chen et al. (2022) revisits this setup while reporting differing observations, they note that models predominantly rely on contextual knowledge in their best-performing settings. They attribute this divergence in findings to two factors. Firstly, the entity substitution approach used by Longpre et al. (2021) potentially reduces the semantic coherence of the perturbed passages. Secondly, Longpre et al. (2021) based their research on single evidence passages, as opposed to Chen et al. (2022), who utilize multiple ones. Recently, with the emergence of really "large" language models such as ChatGPT (Ouyang et al., 2022; OpenAI, 2023a) and Llama (Touvron et al., 2023), *inter alia*, researchers re-examined this issue. Tan et al. (2024) examine how LLMs blend retrieved context with generated knowledge in the ODQA setup, and discover models tend to favor the parametric knowledge, influenced by the greater resemblance of these generated contexts to the input questions and the often incomplete nature of the retrieved information, especially within the scope of conflicting sources.

General. Xie et al. (2023) leverage LLMs to generate conflicting context alongside the memorized knowledge. They find that LLMs are highly receptive to external evidence, even when it conflicts with their parametric, provided that the external knowledge is coherent and convincing. Meanwhile, they also identify a strong confirmation bias (Nickerson, 1998) in LLMs, *i.e.*, the models tend to favor information consistent with their internal memory, even when confronted with conflicting external evidence. Wang et al. (2023g) posit that the desired behaviors when an LLM encounters conflicts should be to pinpoint the conflicts and provide distinct answers. They find while LLMs perform well in identifying the existence of knowledge conflicts, they struggle to determine the specific conflicting segments and produce a response with distinct answers amidst conflicting information. Ying et al. (2023) analyze the robustness of LLMs under conflicts with a focus on two perspectives: factual robustness (the ability to identify correct facts from prompts or memory) and decision style (categorizing LLMs' behavior as intuitive, dependent, or rational-based on cognitive theory). The study finds that LLMs are highly susceptible to misleading prompts, especially in the context of commonsense knowledge. Qian et al. (2023) evaluate the potential interaction between parametric and external knowledge more systematically, cooperating knowledge graph (KG). They reveal that LLMs often deviate from their parametric knowledge when presented with direct conflicts or detailed contextual changes. Xu et al. (2023) study how large language models (LLMs) respond to knowledge conflicts during interactive sessions. Their findings suggest LLMs tend to favor logically structured knowledge, even when it contradicts factual accuracy.

3.3 Mitigation

Mitigation strategies are organized according to their **objectives**, *i.e.*, the desired behaviors we expect from an LLM when it encounters conflicts. Strategies are categorized to 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. *Combining sources* strategies aim for an integrated response leveraging both context and parametric knowledge.

Faithful to Context. *Fine-tuning.* Li et al. (2022a) argues that an LLM should prioritize context for task-relevant information and rely on internal knowledge when the context is unrelated. They name the two properties controllability and robustness. They introduce Knowledge Aware Fine-Tuning (KAFT) to strengthen the two properties by incorporating counterfactual and irrelevant contexts to standard training datasets. Gekhman et al. (2023) introduce TrueTeacher, which focuses on improving factual consistency in summarization by annotating model-generated summaries with LLMs. This approach helps in maintaining faithfulness to the context of the original documents, ensuring that generated summaries remain accurate without being misled by irrelevant or incorrect details. DIAL (Xue et al., 2023) focuses on improving factual consistency in dialogue systems via direct knowledge enhancement and RLFC for aligning responses accurately with provided factual knowledge.

Prompting. Zhou et al. (2023c) explores enhancing LLMs' adherence to context through specialized prompting strategies, specifically opinion-based prompts and counterfactual demonstrations. These techniques are shown to significantly improve LLMs' performance in context-sensitive tasks by ensuring they remain faithful to relevant context, without additional training.

Decoding. Shi et al. (2023a) introduces Context-aware Decoding (CAD) to reduce hallucinations by amplifying the difference in output probabilities with and without context. CAD enhances faithfulness in LLMs by effectively prioritizing relevant context over the model's prior knowledge, especially in tasks with conflicting information.

Knowledge Plug-in. Lee et al. (2022a) proposes Continuously-updated QA (CuQA) for improving LMs' ability to integrate new knowledge. Their approach uses plug-and-play modules to store updated knowledge, ensuring the original model remains unaffected. Unlike traditional continue pre-training or fine-tuning approaches, CuQA can solve knowledge conflicts.

Pre-training. ICLM (Shi et al., 2023b) is a new pre-training method that extends LLMs' ability to handle long and varied contexts across multiple documents. This approach could potentially aid in resolving knowledge conflicts by enabling models to synthesize information from broader contexts, thus improving their understanding and application of relevant knowledge.

Predict Fact Validity. Zhang and Choi (2023) addresses knowledge conflict by introducing fact duration prediction to identify and discard outdated facts in LLMs. This approach improves model performance on tasks like ODQA by ensuring adherence to up-to-date contextual information.

Discriminating Misinformation (Faithful to Memory). *Prompting.* To address misinformation pollution, Pan et al. (2023b) proposes defense strategies such as misinformation detection and vigilant prompting, aiming to enhance the model's ability to remain faithful to factual, parametric information amidst potential misinformation. Similarly, Xu et al. (2023) utilizes a system prompt to remind the LLM to be cautious about potential misinformation and to verify its memorized knowledge before responding. This approach aims to enhance the LLM's ability to maintain faithfulness.

Query Augmentation. Weller et al. (2022) leverages the redundancy of information in large corpora to defend misinformation pollution. Their method involves query augmentation to find a diverse set of less likely poisoned passages, coupled with a confidence method named Confidence from Answer Redundancy (CAR), which compares the predicted answer's consistency across retrieved contexts. This strategy mitigates knowledge conflicts by ensuring the model's faithfulness through cross-verification of answers in multiple sources.

Training Discriminator. Hong et al. (2023) fine-tune a smaller LM as discriminator and combine prompting techniques to develop the model's ability to discriminate between reliable and unreliable information, helping the model remain faithful when confronted with misleading context.

Disentangling Sources. DisentQA (Neeman et al., 2022) trains a model that predicts two types of answers for a given question: one based on contextual knowledge and one on parametric knowledge. Wang et al. (2023g) introduce a method to improve Large Language Models' (LLMs) handling of knowledge conflicts. Their approach is a three-step process designed to help LLMs detect conflicts, accurately identify the conflicting segments, and generate distinct, informed responses based on the conflicting data, aiming for more precise and nuanced model outputs.

Combining Sources. Zhang et al. (2023c) propose COMBO, a framework that pairs compatible generated and retrieved passages to resolve discrepancies. It uses discriminators trained on silver labels to assess passage compatibility, improving ODQA performance by leveraging both LLM-generated (parametric) and external retrieved knowledge.

4 Inter-Context Conflict

4.1 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). Currently, the use of retrieve augment generate (RAG) LLMs has emerged as a novel paradigm in the field. However, the incorporation of retrieved documents as external knowledge introduces a noteworthy concern – the potential for misinformation like fake news within these retrieved documents (Chen et al., 2023b). In the past, there have been instances of AI being utilized for the generation of misinformation (Zhou et al., 2023b; Vergho et al., 2024; Weidinger et al., 2021). With the formidable generative capabilities of LLMs, this issue has been further exacerbated, contributing to a significant surge in 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 (Chen et al., 2021; Liska et al., 2022; Zhang and Choi, 2021; Kasai et al., 2022).

4.2 Analysis

Performance Impact. Previous research has empirically demonstrated that the performance of a pre-trained 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 a recent study, Pan et al. (2023a) introduced a misinformation attack strategy involving the creation of fabricated versions of Wikipedia articles, which are subsequently inserted into the authentic Wikipedia corpus. Their research findings revealed that existing language models are susceptible to misinformation attacks, irrespective of whether the fake articles are manually crafted or generated by models. Notably, when the retrieval dataset includes 4% of misinformation, the model's Exact Match (EM) performance drops by approximately 10%. To gain a deeper understanding of how LLMs behave when encountering contradictory contexts, Chen et al. (2022) primarily conducted experiments using Fusion-in-Decoder on the NQ-Open (Kwiatkowski et al., 2019) and TriviaQA (Joshi et al., 2017). They found that contradictions within knowledge sources have a marginal impact on models' confidence and models tend to prioritize context that is more relevant to the query and context that contains answers consistent with the model's parametric knowledge. Xie et al. (2023) conducted experiments on both closed-source LLMs and open-source LLMs in POPQA (Mallen et al., 2022) and StrategyQA (Geva et al., 2021). The results obtained were in line with those of Chen et al. (2022), indicating that LLMs exhibit a significant bias to evidence that aligns with the model's parametric memory. They also found that LLMs tend to place stronger emphasis on facts related to more popular entities and answers supported by more documents in the context, and are highly sensitive to the order in which information is presented.

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) examines 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 LLaMAv2 in identifying contradictory documents within news articles (Hermann et al., 2015), stories (Kočiskỳ et al., 2018), and wikipedia (Merity et al., 2016). The authors found that, even for GPT-4, the average detection accuracy remains at around 70%. The study also revealed that LLMs encounter particular challenges when dealing with specific types of contradiction, notably those related to subjective emotions or perspectives and the length of the documents and the range of self-contradictions have a slight impact on the detection performance.

4.3 Mitigation

Eliminating Conflict. *Specialized Models.* Hsu et al. (2021) develop a model named Pairwise Contradiction Neural Network (PCNN), which generates contradiction probabilities based on sentence representations obtained through fine-tuned Sentence-BERT. They conducted on Wikipedia various proportions of misinformation to demonstrate the effectiveness of their approach. Pielka et al. (2022) discovered that XLM-RoBERTa struggles to effectively grasp the syntactic and semantic features that contribute to incorrect contradiction detection. They suggested incorporating linguistic knowledge into the learning process. Wu et al. (2022) developed a novel approach that integrates topological representations of text into deep learning models. They performed experiments using three widely-used models including BERT, ESIM, and CBOW on MultiNLI dataset (Williams et al., 2017). The results provide compelling evidence that their approach is effective.

General Models. Chern et al. (2023) proposed a fact-checking framework that integrates LLMs including GPT4, GPT3.5 and FLAN-T5-XXL with various tools, such as Google Search, Google Scholar, code interpreters, and Python, for detecting factual errors in texts. The authors conducted experiments using RoSE (Liu et al., 2022) and Fact-Prompts. Cheung and Lam (2023) combined the search engine with Llama to predict the veracity of claims. They conducted experiments using two datasets, RAWFC and LIAR. Leite et al. (2023) employ LLMs to produce weak labels associated with 18 credibility signals for the input text and aggregate these labels through weak supervision techniques to make predictions regarding the veracity of the input. Effectiveness of their method on the FA-KES and EUvsDisinfo datasets.

Improving Robustness. *Training Approach.* Hong et al. (2023) presents a novel fine-tuning method that involves training a discriminator alongside the decoder using the same encoder of FiD. Additionally, the author introduces two other methods to improve the robustness of the model including prompting GPT-3 to identify perturbed documents before generating responses and integrating the discriminator's output into the prompt for GPT-3. They conduct experiments on NQ-Open with entity replacement operations (Longpre et al., 2021) and their experimental results indicate that the fine-tuning method yields the most promising results.

Query Augmentation. Weller et al. (2022) first prompt GPT-3 to generate new questions based on the original question and then measure the confidence for each query and its corresponding retrieved passages. This confidence is used to determine whether to use the original question's prediction or to opt for a majority vote based on the predictions obtained from the augmented questions that exhibit a high degree of confidence. They verify the effectiveness of their method on Natural Questions and TriviaQA.

5 Conclusion

Through this survey, we have highlighted the importance of investigating knowledge conflicts, shedding light on their categorization, causes, behavioral patterns, and potential mitigation approaches. We have observed that existing studies have started to recognize the significance of knowledge conflicts but lack a systematic exploration dedicated solely to this topic. Our survey aims to fill this gap by providing a comprehensive review, synthesizing relevant literature, and offering insights into potential avenues for further research and resolution.

Limitations

Given the rapid growth of research in the field of knowledge conflict and the abundance of research literature, it is inevitable that we may overlook some most recent or less relative findings. However, we have included all the most essential materials in our survey.

Ethics Statement

No Consideration.

References

Abdalghani Abujabal, Rishiraj Saha Roy, Mohamed Yahya, and Gerhard Weikum. 2018. Comqa: A community-sourced dataset for complex factoid question answering with paraphrase clusters. *arXiv preprint arXiv:1809.09528*.

Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. 2021. Explanations for commonsenseqa: New dataset and models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3050–3065.

Ayush Agrawal, Lester Mackey, and Adam Tauman Kalai. 2023. Do language models know when they're hallucinating references? *arXiv preprint arXiv:2305.18248*.

Ekin Akyürek, Tolga Bolukbasi, Frederick Liu, Binbin Xiong, Ian Tenney, Jacob Andreas, and Kelvin Guu. 2022. Towards tracing knowledge in language models back to the training data. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2429–2446.

Navid Ayoobi, Sadat Shahriar, and Arjun Mukherjee. 2023. The looming threat of fake and llm-generated linkedin profiles: Challenges and opportunities for detection and prevention. In *Proceedings of the 34th ACM Conference on Hypertext and Social Media*, pages 1–10.

Clark Barrett, Brad Boyd, Elie Bursztein, Nicholas Carlini, Brad Chen, Jihye Choi, Amrita Roy Chowdhury, Mihai Christodorescu, Anupam Datta, Soheil Feizi, et al. 2023. Identifying and mitigating the security risks of generative ai. *Foundations and Trends® in Privacy and Security*, 6(1):1–52.

Henning Bartsch, Ole Jorgensen, Domenic Rosati, Jason Hoelscher-Obermaier, and Jacob Pfau. 2023. Self-consistency of large language models under ambiguity. *arXiv preprint arXiv:2310.13439*.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, Gillian Hadfield, et al. 2023. Managing ai risks in an era of rapid progress. *arXiv preprint arXiv:2310.17688*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Nicholas Carlini, Matthew Jagielski, Christopher A Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. 2023. Poisoning web-scale training datasets is practical. *arXiv preprint arXiv:2302.10149*.

Tyler A Chang and Benjamin K Bergen. 2023. Language model behavior: A comprehensive survey. *arXiv preprint arXiv:2303.11504*.

Chi Cheang, Hou Chan, Derek Wong, Xuebo Liu, Zhaocong Li, Yanming Sun, Shudong Liu, and Lidia Chao. 2023. Can lms generalize to future data? an empirical analysis on text summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16205–16217.

Canyu Chen and Kai Shu. 2023a. Can llm-generated misinformation be detected? In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Canyu Chen and Kai Shu. 2023b. Combating misinformation in the age of llms: Opportunities and challenges. *arXiv preprint arXiv:2311.05656*.

Hung-Ting Chen, Michael JQ Zhang, and Eunsol Choi. 2022. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. *arXiv preprint arXiv:2210.13701*.

Jiangjie Chen, Wei Shi, Ziquan Fu, Sijie Cheng, Lei Li, and Yanghua Xiao. 2023a. Say what you mean! large language models speak too positively about negative commonsense knowledge. *arXiv preprint arXiv:2305.05976*.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2023b. Benchmarking large language models in retrieval-augmented generation. *arXiv preprint arXiv:2309.01431*.

Liang Chen, Yang Deng, Yatao Bian, Zeyu Qin, Bingzhe Wu, Tat-Seng Chua, and Kam-Fai Wong. 2023c. Beyond factuality: A comprehensive evaluation of large language models as knowledge generators. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6325–6341.

Wenhu Chen, Xinyi Wang, and William Yang Wang. 2021. A dataset for answering time-sensitive questions. *arXiv preprint arXiv:2108.06314*.

I Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, et al. 2023. Factool: Factuality detection in generative ai–a tool augmented framework for multi-task and multi-domain scenarios. *arXiv preprint arXiv:2307.13528*.

Tsun-Hin Cheung and Kin-Man Lam. 2023. Factllama: Optimizing instruction-following language models with external knowledge for automated fact-checking. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 846–853. IEEE.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936.

Leyang Cui, Sijie Cheng, Yu Wu, and Yue Zhang. 2020. Does bert solve commonsense task via commonsense knowledge. *arXiv preprint arXiv:2008.03945*, 4.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. 2021. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3366–3385.

Bhuwan Dhingra, Jeremy R Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*.

Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Zhifang Sui, and Lei Li. 2023. Statistical knowledge assessment for large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2022a. e-care: a new dataset for exploring explainable causal reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 432–446.

Yibing Du, Antoine Bosselut, and Christopher D Manning. 2022b. Synthetic disinformation attacks on automated fact verification systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10581–10589.

Nouha Dziri, Andrea Madotto, Osmar Zaiane, and Avishek Joey Bose. 2021. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. *arXiv preprint arXiv:2104.08455*.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Amir Feder, Abhilasha Ravichander, Marius Mosbach, Yonatan Belinkov, Hinrich Schütze, and Yoav Goldberg. 2022. Measuring causal effects of data statistics on language model'sfactual'predictions. *arXiv preprint arXiv:2207.14251*.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. 2018. T-rex: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*.

Zhangyin Feng, Weitao Ma, Weijiang Yu, Lei Huang, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. Trends in integration of knowledge and large language models: A survey and taxonomy of methods, benchmarks, and applications. *arXiv preprint arXiv:2311.05876*.

Emilio Ferrara. 2023. Genai against humanity: Nefarious applications of generative artificial intelligence and large language models. *arXiv preprint arXiv:2310.00737*.

Yi R Fung, Kung-Hsiang Huang, Preslav Nakov, and Heng Ji. 2022. The battlefront of combating misinformation and coping with media bias. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4790–4791.

Wee Chung Gan and Hwee Tou Ng. 2019. Improving the robustness of question answering systems to question paraphrasing. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 6065–6075.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, et al. 2023a. Rarr: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023b. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Zorik Gekhman, Jonathan Herzig, Roee Aharoni, Chen Elkind, and Idan Szpektor. 2023. Trueteacher: Learning factual consistency evaluation with large language models. *arXiv preprint arXiv:2305.11171*.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Josh A Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative language models and automated influence operations: Emerging threats and potential mitigations. *arXiv preprint arXiv:2301.04246*.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. More than you've asked for: A comprehensive analysis of novel prompt injection threats to application-integrated large language models. *arXiv e-prints*, pages arXiv–2302.

Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, et al. 2023. Studying large language model generalization with influence functions. *arXiv preprint arXiv:2308.03296*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Peter Hase, Mona Diab, Asli Celikyilmaz, Xian Li, Zornitsa Kozareva, Veselin Stoyanov, Mohit Bansal, and Srinivasan Iyer. 2023. Methods for measuring, updating, and visualizing factual beliefs in language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2706–2723.

Tianxing He, Jun Liu, Kyunghyun Cho, Myle Ott, Bing Liu, James Glass, and Fuchun Peng. 2021. Analyzing the forgetting problem in pretrain-finetuning of open-domain dialogue response models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1121–1133.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Giwon Hong, Jeonghwan Kim, Junmo Kang, Sung-Hyon Myaeng, and Joyce Jiyoung Whang. 2023. Discern and answer: Mitigating the impact of misinformation in retrieval-augmented models with discriminators. *arXiv preprint arXiv:2305.01579*.

Cheng Hsu, Cheng-Te Li, Diego Saez-Trumper, and Yi-Zhan Hsu. 2021. Wikicontradiction: Detecting self-contradiction articles on wikipedia. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 427–436. IEEE.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2022. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*.

Joel Jang, Seonghyeon Ye, Changho Lee, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, and Minjoon Seo. 2022a. Temporalwiki: A lifelong benchmark for training and evaluating ever-evolving language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6237–6250.

Joel Jang, Seonghyeon Ye, Sohee Yang, Joongbo Shin, Janghoon Han, KIM Gyeonghun, Stanley Jungkyu Choi, and Minjoon Seo. 2021. Towards continual knowledge learning of language models. In *International Conference on Learning Representations*.

Myeongjun Jang, Deuk Sin Kwon, and Thomas Lukasiewicz. 2022b. Becel: Benchmark for consistency evaluation of language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3680–3696.

Myeongjun Erik Jang and Thomas Lukasiewicz. 2023. Improving language models meaning understanding and consistency by learning conceptual roles from dictionary. *arXiv preprint arXiv:2310.15541*.

Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks VS Lakshmanan. 2020. Automatic detection of machine generated text: A critical survey. *arXiv preprint arXiv:2011.01314*.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Bohan Jiang, Zhen Tan, Ayushi Nirmal, and Huan Liu. 2023. Disinformation detection: An evolving challenge in the age of llms. *arXiv preprint arXiv:2309.15847*.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pages 15696–15707. PMLR.

Cheongwoong Kang and Jaesik Choi. 2023. Impact of co-occurrence on factual knowledge of large language models. *arXiv preprint arXiv:2310.08256*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, and Kentaro Inui. 2022. Realtime qa: What's the answer right now? *arXiv preprint arXiv:2207.13332*.

Nora Kassner, Oyvind Tafjord, Hinrich Schütze, and Peter Clark. 2021. Beliefbank: Adding memory to a pre-trained language model for a systematic notion of belief. *arXiv preprint arXiv:2109.14723*.

Celeste Kidd and Abeba Birhane. 2023. How ai can distort human beliefs. *Science*, 380(6651):1222–1223.

Miyoung Ko, Ingyu Seong, Hwaran Lee, Joonsuk Park, Minsuk Chang, and Minjoon Seo. 2022. Claimdiff: Comparing and contrasting claims on contentious issues. *arXiv preprint arXiv:2205.12221*.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Srijan Kumar and Neil Shah. 2018. False information on web and social media: A survey. *arXiv preprint arXiv:1804.08559*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Tsz Kin Lam, Eva Hasler, and Felix Hieber. 2022. Analyzing the use of influence functions for instance-specific data filtering in neural machine translation. *arXiv preprint arXiv:2210.13281*.

Angeliki Lazaridou, Elena Gribovskaya, Wojciech Stokowiec, and Nikolai Grigorev. 2022. Internet-augmented language models through few-shot prompting for open-domain question answering. *arXiv preprint arXiv:2203.05115*.

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. 2021. Mind the gap: Assessing temporal generalization in neural language models. *Advances in Neural Information Processing Systems*, 34:29348–29363.

Kyungjae Lee, Wookje Han, Seung-won Hwang, Hwaran Lee, Joonsuk Park, and Sang-Woo Lee.

2022a. Plug-and-play adaptation for continuously-updated qa. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 438–447.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale N Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2022b. Factuality enhanced language models for open-ended text generation. *Advances in Neural Information Processing Systems*, 35:34586–34599.

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

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2022a. Large language models with controllable working memory. *arXiv preprint arXiv:2211.05110*.

Jierui Li, Vipul Raheja, and Dhruv Kumar. 2023a. Contradoc: Understanding self-contradictions in documents with large language models. *arXiv preprint arXiv:2311.09182*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023b. Inference-time intervention: Eliciting truthful answers from a language model. *arXiv preprint arXiv:2306.03341*.

Shaobo Li, Xiaoguang Li, Lifeng Shang, Zhenhua Dong, Chengjie Sun, Bingquan Liu, Zhenzhou Ji, Xin Jiang, and Qun Liu. 2022b. How pre-trained language models capture factual knowledge? a causal-inspired analysis. *arXiv preprint arXiv:2203.16747*.

Xiang Lisa Li, Vaishnavi Shrivastava, Siyan Li, Tatsunori Hashimoto, and Percy Liang. 2023c. Benchmarking and improving generator-validator consistency of language models. *arXiv preprint arXiv:2310.01846*.

Zhoubo Li, Ningyu Zhang, Yunzhi Yao, Mengru Wang, Xi Chen, and Huajun Chen. 2023d. Unveiling the pitfalls of knowledge editing for large language models. *arXiv preprint arXiv:2310.02129*.

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

Adam Liska, Tomas Kocisky, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, D'Autume Cyprien De Masson, Tim Scholtes, Manzil Zaheer, Susannah Young, et al. 2022. Streamingqa: A benchmark for adaptation to new knowledge over time in question answering models. In *International Conference on Machine Learning*, pages 13604–13622. PMLR.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, and Yang Liu. 2023b. Prompt injection attack against llm-integrated applications. *arXiv preprint arXiv:2306.05499*.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063.

Kelvin Luu, Daniel Khashabi, Suchin Gururangan, Karishma Mandyam, and Noah A Smith. 2021. Time waits for no one! analysis and challenges of temporal misalignment. *arXiv preprint arXiv:2111.07408*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Hannaneh Hajishirzi, and Daniel Khashabi. 2022. When not to trust language models: Investigating effectiveness and limitations of parametric and non-parametric memories. *arXiv preprint arXiv:2212.10511*, 7.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

Katerina Margatina, Shuai Wang, Yogarshi Vyas, Neha Anna John, Yassine Benajiba, and Miguel Ballesteros. 2023. Dynamic benchmarking of masked language models on temporal concept drift with multiple views. *arXiv preprint arXiv:2302.12297*.

Luca Massarelli, Fabio Petroni, Aleksandra Piktus, Myle Ott, Tim Rocktäschel, Vassilis Plachouras, Fabrizio Silvestri, and Sebastian Riedel. 2019. How decoding strategies affect the verifiability of generated text. *arXiv preprint arXiv:1911.03587*.

Priyanka Meel and Dinesh Kumar Vishwakarma. 2020. Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities. *Expert Systems with Applications*, 153:112986.

Filippo Menczer, David Crandall, Yong-Yeol Ahn, and Apu Kapadia. 2023. Addressing the harms of ai-generated inauthentic content. *Nature Machine Intelligence*, 5(7):679–680.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2021. Fast model editing at scale. *arXiv preprint arXiv:2110.11309*.

Eric Mitchell, Joseph J Noh, Siyan Li, William S Armstrong, Ananth Agarwal, Patrick Liu, Chelsea Finn, and Christopher D Manning. 2022. Enhancing self-consistency and performance of pre-trained language models through natural language inference. *arXiv preprint arXiv:2211.11875*.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2023. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. *arXiv preprint arXiv:2305.15852*.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2022. Disentqa: Disentangling parametric and contextual knowledge with counterfactual question answering. *arXiv preprint arXiv:2211.05655*.

Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220.

Xenia Ohmer, Elia Bruni, and Dieuwke Hupkes. 2023. Separating form and meaning: Using self-consistency to quantify task understanding across multiple senses. *CoRR*.

Yasumasa Onoe, Michael JQ Zhang, Shankar Padmanabhan, Greg Durrett, and Eunsol Choi. 2023. Can lms learn new entities from descriptions? challenges in propagating injected knowledge. *arXiv preprint arXiv:2305.01651*.

OpenAI. 2023a. Chatgpt.

OpenAI. 2023b. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Liangming Pan, Wenhu Chen, Min-Yen Kan, and William Yang Wang. 2023a. Attacking open-domain question answering by injecting misinformation. *IJCNLP-AACL. ACL*.

Xiaoman Pan, Wenlin Yao, Hongming Zhang, Dian Yu, Dong Yu, and Jianshu Chen. 2022. Knowledge-in-context: Towards knowledgeable semi-parametric language models. In *The Eleventh International Conference on Learning Representations*.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Yang Wang. 2023b. On the risk of misinformation pollution with large language models. *arXiv preprint arXiv:2305.13661*.

Jungsoo Park, Sewon Min, Jaewoo Kang, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2021. Faviq: Fact verification from information-seeking questions. *arXiv preprint arXiv:2107.02153*.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. 2022. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473.

Maren Pielka, Felix Rode, Lisa Pucknat, Tobias Deußer, and Rafet Sifa. 2022. A linguistic investigation of machine learning based contradiction detection models: an empirical analysis and future perspectives. In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1649–1653. IEEE.

Yuval Pinter and Michael Elhadad. 2023. Emptying the ocean with a spoon: Should we edit models? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15164–15172.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-lingual consistency of factual knowledge in multilingual language models. *arXiv preprint arXiv:2310.10378*.

Cheng Qian, Xinran Zhao, and Sherry Tongshuang Wu. 2023. " merge conflicts!" exploring the impacts of external distractors to parametric knowledge graphs. *arXiv preprint arXiv:2309.08594*.

Ella Rabinovich, Samuel Ackerman, Orna Raz, Eitan Farchi, and Ateret Anaby-Tavor. 2023. Predicting question-answering performance of large language models through semantic consistency. *arXiv preprint arXiv:2311.01152*.

Harsh Raj, Vipul Gupta, Domenic Rosati, and Subhabrata Majumdar. 2023. Semantic consistency for assuring reliability of large language models. *arXiv preprint arXiv:2308.09138*.

Harsh Raj, Domenic Rosati, and Subhabrata Majumdar. 2022. Measuring reliability of large language models through semantic consistency. *arXiv preprint arXiv:2211.05853*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*.

Or Sharir, Barak Peleg, and Yoav Shoham. 2020. The cost of training nlp models: A concise overview. *arXiv preprint arXiv:2004.08900*.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen-tau Yih. 2023a. Trusting your evidence: Hallucinate less with context-aware decoding. *arXiv preprint arXiv:2305.14739*.

Weijia Shi, Sewon Min, Maria Lomeli, Chunting Zhou, Margaret Li, Victoria Lin, Noah A Smith, Luke Zettlemoyer, Scott Yih, and Mike Lewis. 2023b. In-context pretraining: Language modeling beyond document boundaries. *arXiv preprint arXiv:2310.10638*.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023c. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2022. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*.

Anton Sinitsin, Vsevolod Plokhotnyuk, Dmitry Pyrkin, Sergei Popov, and Artem Babenko. 2019. Editable neural networks. In *International Conference on Learning Representations*.

Craig S. Smith. 2023. What large models cost you – there is no free ai lunch.

Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Hal Daumé III, Jesse Dodge, Ellie Evans, Sara Hooker, et al. 2023. Evaluating the social impact of generative ai systems in systems and society. *arXiv preprint arXiv:2306.05949*.

Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. 2023. Ai model gpt-3 (dis) informs us better than humans. *arXiv preprint arXiv:2301.11924*.

Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, and Xueqi Cheng. 2024. Blinded by generated contexts: How language models merge generated and retrieved contexts for open-domain qa? *arXiv preprint arXiv:2401.11911*.

Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. 2023. The science of detecting llm-generated texts. *arXiv preprint arXiv:2303.07205*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *arXiv preprint arXiv:2305.04388*.

Tyler Vergho, Jean-Francois Godbout, Reihaneh Rabbany, and Kellin Pelrine. 2024. Comparing gpt-4 and open-source language models in misinformation mitigation. *arXiv preprint arXiv:2401.06920*.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, et al. 2023. Freshllms: Refreshing large language models with search engine augmentation. *arXiv preprint arXiv:2310.03214*.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, Yidong Wang, Linyi Yang, Jindong Wang, Xing Xie, Zheng Zhang, and Yue Zhang. 2023a. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity.

Cunxiang Wang, Zhikun Xu, Qipeng Guo, Xiangkun Hu, Xuefeng Bai, Zheng Zhang, and Yue Zhang. 2023b. Exploiting Abstract Meaning Representation for open-domain question answering. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2083–2096, Toronto, Canada. Association for Computational Linguistics.

Cunxiang Wang, Haofei Yu, and Yue Zhang. 2023c. RFiD: Towards rational fusion-in-decoder for open-domain question answering. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2473–2481, Toronto, Canada. Association for Computational Linguistics.

Fei Wang, Wenjie Mo, Yiwei Wang, Wenxuan Zhou, and Muhao Chen. 2023d. A causal view of entity bias in (large) language models. *arXiv preprint arXiv:2305.14695*.

Jiaan Wang, Yunlong Liang, Zengkui Sun, Yuxuan Cao, and Jiarong Xu. 2023e. Cross-lingual knowledge editing in large language models. *arXiv preprint arXiv:2309.08952*.

Liyuan Wang, Xingxing Zhang, Qian Li, Mingtian Zhang, Hang Su, Jun Zhu, and Yi Zhong. 2023f. Incorporating neuro-inspired adaptability for continual learning in artificial intelligence. *Nature Machine Intelligence*, pages 1–13.

Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2023g. Resolving knowledge conflicts in large language models. *arXiv preprint arXiv:2310.00935*.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V Le. 2023. Simple synthetic data reduces sycophancy in large language models. *arXiv preprint arXiv:2308.03958*.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, et al. 2023. Sociotechnical safety evaluation of generative ai systems. *arXiv preprint arXiv:2310.11986*.

Orion Weller, Aleem Khan, Nathaniel Weir, Dawn Lawrie, and Benjamin Van Durme. 2022. Defending against misinformation attacks in open-domain question answering. *arXiv preprint arXiv:2212.10002*.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Xiangcheng Wu, Xi Niu, and Ruhani Rahman. 2022. Topological analysis of contradictions in text. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2478–2483.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Unraveling the behavior of large language models in knowledge conflicts. *arXiv preprint arXiv:2305.13300*.

Nan Xu, Fei Wang, Bangzheng Li, Mingtao Dong, and Muhao Chen. 2022. Does your model classify entities reasonably? diagnosing and mitigating spurious correlations in entity typing. *arXiv preprint arXiv:2205.12640*.

Rongwu Xu, Brian S Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. 2023. The earth is flat because...: Investigating llms' belief towards misinformation via persuasive conversation. *arXiv preprint arXiv:2312.09085*.

Boyang Xue, Weichao Wang, Hongru Wang, Fei Mi, Rui Wang, Yasheng Wang, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2023. Improving factual consistency for knowledge-grounded dialogue systems via knowledge enhancement and alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7829–7844.

Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2023. Editing large language models: Problems, methods, and opportunities. *arXiv preprint arXiv:2305.13172*.

Jingwei Yi, Yueqi Xie, Bin Zhu, Keegan Hines, Emre Kiciman, Guangzhong Sun, Xing Xie, and Fangzhao Wu. 2023. Benchmarking and defending against indirect prompt injection attacks on large language models. *arXiv preprint arXiv:2312.14197*.

Jiahao Ying, Yixin Cao, Kai Xiong, Yidong He, Long Cui, and Yongbin Liu. 2023. Intuitive or dependent? investigating llms' robustness to conflicting prompts. *arXiv preprint arXiv:2309.17415*.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2022. Generate rather than retrieve: Large language models are strong context generators. *arXiv preprint arXiv:2209.10063*.

Jiaxin Zhang, Zhuohang Li, Kamalika Das, Bradley A Malin, and Sricharan Kumar. 2023a. Sac^3: Reliable hallucination detection in black-box language models via semantic-aware cross-check consistency. *arXiv preprint arXiv:2311.01740*.

Michael JQ Zhang and Eunsol Choi. 2021. Situatedqa: Incorporating extra-linguistic contexts into qa. *arXiv preprint arXiv:2109.06157*.

Michael JQ Zhang and Eunsol Choi. 2023. Mitigating temporal misalignment by discarding outdated facts. *arXiv preprint arXiv:2305.14824*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023b. Siren's song in the ai ocean: A survey on hallucination in large language models.

Yunxiang Zhang, Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. 2023c. Merging generated and retrieved knowledge for open-domain qa. *arXiv preprint arXiv:2310.14393*.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2023a. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*.

Yukun Zhao, Lingyong Yan, Weiwei Sun, Guoliang Xing, Chong Meng, Shuaiqiang Wang, Zhicong Cheng, Zhaochun Ren, and Dawei Yin. 2023b. Knowing what llms do not know: A simple yet effective self-detection method. *arXiv preprint arXiv:2310.17918*.

Chujie Zheng, Jinfeng Zhou, Yinhe Zheng, Libiao Peng, Zhen Guo, Wenquan Wu, Zhengyu Niu, Hua Wu, and Minlie Huang. 2022. Cdconv: A benchmark for contradiction detection in chinese conversations. *arXiv preprint arXiv:2210.08511*.

Zexuan Zhong, Zhengxuan Wu, Christopher D Manning, Christopher Potts, and Danqi Chen. 2023. Mquake: Assessing knowledge editing in language models via multi-hop questions. *arXiv preprint arXiv:2305.14795*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023a. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*.

Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023b. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–20.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023c. Context-faithful prompting for large language models. *arXiv preprint arXiv:2303.11315*.

Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. 2023. Toolqa: A dataset for llm question answering with external tools. *arXiv preprint arXiv:2306.13304*.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)*, 51(2):1–36.

A Intra-Memory Conflict

Recently, LLMs have gained widespread utilization across various domains, especially in knowledge-intensive question-answering systems (Gao et al., 2023b; Yu et al., 2022; Petroni et al., 2019; Chen et al., 2023c). The deployment of LLMs securely and dependably hinges upon ensuring the consistency of their outputs when presented with expressions conveying similar meanings or intents. However, a significant challenge emerges in the form of intra-memory conflict within LLMs. Intra-memory conflict pertains to the phenomenon wherein the

Datasets	Approach[1]	Base[2]	Size	Conflict
Xie et al. (2023)	Gen	PopQA (2023), STRATEGYQA ((Geva et al., 2021))	20,091	CM[3]
KC (2023g)	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
Pan et al. (2023a)	Gen,Sub	SQuAD 1.1 (2016)	52,189	IC
CONTRADOC (2023a)	Gen	CNN-DailyMail (2015), NarrativeQA (2018) WikiText (2016)	449	IC
ClaimDiff (2022)	Hum	N/A	2,941	IC
WikiContradiction (2021)	Hum	Wikipedia	2,210	IC
PARAREL (2021)	Hum	T-REx (2018)	328	IM

1. Approach refers to how the conflicts are crafted, including entity-level substitution (Sub), generative approaches employing an LLM (Gen), human annotation (Hum).
2. Base refers to the base dataset(s) that serve as the foundation for generating conflicts, if applicable.
3. ⚠ For CM datasets, conflicts are derived from a *certain* model's parametric knowledge, which can vary between models. Therefore, application requires selecting a subset of the dataset that aligns with the tested model's knowledge.

Table 1: Datasets on evaluating a large language model's behavior when encounter knowledge conflicts.

model's parameters encompass multiple, potentially conflicting versions of knowledge. As a result, the model may exhibit unpredictable behavior and different outputs for inputs that, while syntactically varied, convey the same semantic information (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 substantially diminishes the practicality and efficacy of LLMs.

A.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., 2023d). These factors respectively pertain to the training phase, the inference phase, and subsequent knowledge refinement.

Bias in Training Corpora. Recent research demonstrates that the primary phase for knowledge acquisition in LLMs predominantly occurs in the pre-training stage (Zhou et al., 2023a; Kaddour et al., 2023; Naveed et al., 2023; Akyürek et al., 2022; Singhal et al., 2022). Pre-training data is primarily crawled from the internet, which exhibits a diverse range of data quality, potentially including inaccurate or misleading information (Bender et al., 2021; Weidinger et al., 2021). When LLMs are trained on data containing incorrect knowledge, they may memorize and inadvertently amplify these inaccuracies (Lin et al., 2022; Elazar et al., 2022; Lam et al., 2022; Grosse et al., 2023), leading to a situation where conflicting knowledge coexists within the parameters of LLMs concurrently.

Moreover, prior works indicate that LLMs may possess a propensity for encoding superficial associations prevalent within their training data, as opposed to genuinely comprehending the underlying knowledge contained therein (Li et al., 2022b; Kang and Choi, 2023; Zhao et al., 2023a; Kandpal et al., 2023). Consequently, when the training dataset exhibits a bias towards spurious correlations, it can result in the LLMs displaying a propensity to generate predetermined responses rooted in those spurious correlations. Due to the dependency of spurious correlations, LLMs may provide divergent answers when presented with prompts exhibiting distinct syntactic structures but conveying equivalent semantic meaning, thereby leading to instances of intra-memory conflicts.

Decoding Strategy. The direct output of LLMs is a probability distribution representing the possible next tokens. Sampling is a crucial step in determining the generated content from this distribution. Currently, there are various proposed sampling techniques, including greedy sampling, top-p sampling, top-k sampling, and others (Jawahar et al., 2020; Massarelli et al., 2019). These techniques can be categorized into two main groups: deterministic sampling and stochastic sampling. Stochastic sampling stands as the prevailing decoding strategy employed by LLMs (Fan et al., 2018; Holtzman et al., 2019). However, the stochastic nature of sampling introduces uncertainty into the generated content. Furthermore, due to the intrinsic left-to-right generation pattern of LLMs, the selection of the sampling token can wield a significant influence over the content of subsequent generations. The use of stochastic sampling may lead LLMs to produce entirely different content, even when provided with the same context, causing intra-memory conflict (Lee et al., 2022b; Huang et al., 2023; Dziri et al., 2021).

Knowledge Editing. With the dramatic increase of model parameters, fine-tuning LLMs become increasingly challenging and resource-intensive. In response to this challenge, researchers have turned to knowledge editing techniques as a means to efficiently modify the small scope of knowledge learned of LLMs (Meng et al., 2022; Ilharco et al., 2022; Zhong et al., 2023). Ensuring the consistency of modifications poses a significant challenge. Due to the potential defects inherent in the editing method, the modified knowledge cannot be generalized effectively. Consequently, the model exhibits variations in its responses across different contexts. It may adapt its knowledge to specific situations while maintaining consistency in others (Li et al., 2023d; Yao et al., 2023). Intra-memory conflict is primarily considered a side effect in the context of knowledge editing.

A.2 Analysis

Self-Inconsistency. Elazar et al. (2021) developed a method for assessing the knowledge consistency of a model, focusing specifically on knowledge triples. The authors primarily conducted experiments using BERT, RoBERTa, and ALBERT, and their findings indicate that these models exhibit poor consistency, with guaranteed accuracy ranging from only 50% to 60% on the test data. Hase et al. (2023) employed the same indicators of Elazar et al. (2021), but they utilized a more diverse dataset. Their study also revealed that the consistency of RoBERTa-base and BART-base within the paraphrase context was lacking. Zhao et al. (2023b) first reformulated the questions and then assessed the consistency of the LLM's responses to these reformulated questions. The findings of their research revealed that even GPT-4 exhibits a notable inconsistency rate of 13% when applied to Commonsense Question-Answering tasks. They further found that LLMs are more likely to produce inconsistencies in the face of uncommon knowledge. Dong et al. (2023) conducted experiments on 20 open-source LLMs and found that all of these models exhibit strong inconsistencies. The authors also found that fine-tuning LLMs on data collected from a more knowledgeable model could augment its knowledge. Li et al. (2023c) explored an additional aspect of inconsistency that LLMs can give an initial answer to a question, but it may subsequently contradict that answer when asked if it's correct. The authors conducted experiments focusing on Close-Book Question Answering and revealed that GPT-4 is consistent on 95% of cases, while Alpaca-30B only displays consistency in 50% of cases.

To further analyze the inconsistency exhibited by LLMs, a study conducted by Li et al. (2022b) revealed that encoder-based models tend to generate missing factual words more relying on positionally close and highly co-occurring words, rather than knowledge-dependent words. This phenomenon arises due to these models' tendency to overlearn inappropriate associations from the training dataset. Kang and Choi (2023) demonstrated that LLMs are prone to co-occurrence bias, where they favor frequently co-occurring words over the correct answer. Furthermore, their research highlighted that LLMs face challenges in recalling facts in cases where the subject and object rarely appear together in the pre-training dataset, even though these facts are encountered during fine-tuning.

Latent Representation of Knowledge. Contemporary large language models all employ a multi-layer transformer structure, giving rise to a unique form of inter-memory conflict, *i.e.*the presence of distinct knowledge representations across different layers of a model. In the past, numerous researchers have proposed that a language model would store low-level information at a shallow level, and semantic information at a high level (Tenney et al., 2019; Rogers et al., 2021; Jawahar et al., 2019; Cui et al., 2020). Chuang et al. (2023) explored this aspect within the context of LLMs and discovered that factual knowledge in LLMs is typically concentrated within specific transformer layers and different layers of inconsistent knowledge. Moreover, Li et al. (2023b) discovered that the correct knowledge is indeed stored within the parameters of the large model, but it may not be accurately expressed during the generation process. The authors conducted two experiments on the same LLaMa 7B, one focused on the generation accuracy, and the other utilizing a knowledge probe to examine the knowledge containment. The results of these experiments revealed a substantial 40% disparity between the knowledge probe accuracy and the generation accuracy.

Cross-lingual Inconsistency. The meaning of true knowledge is not influenced by surface form (Ohmer et al., 2023). Knowledge held by LLMs should also possess this characteristic. However, unlike humans, LLMs maintain distinct sets of knowledge for various languages (Ji et al., 2023), leading to potential inconsistencies in their knowledge across different languages. Wang et al. (2023e) analyzed LLMs' knowledge expressed in one language after implementing knowledge editing on this knowledge expressed in another language. Their findings suggest that LLMs face difficulties when attempting to extend the edited knowledge to other languages and exhibit inconsistent behaviors. These findings indicate that knowledge related to different languages is stored separately within the model parameters, introducing a risk of intra-memory conflict of the model. Qi et al. (2023) conducted a more direct study. They propose a metric named RankC for evaluating the cross-lingual consistency of factual knowledge of LLMs. They employed this metric for analyzing multiple models and revealed that the knowledge learned by LLMs is not language-agnostic. Instead, a strong language dependence exists in the knowledge of LLMs. Additionally, they found that increasing the model size does not lead to an improvement in cross-lingual consistency.

A.3 Mitigation

A.3.1 Improving Consistency

Fine-tuning. Elazar et al. (2021) introduces a new benchmark called PARALLEL, which consists of 328 paraphrases that describe 38 binary relationships. The author proposed the consistency loss function and leveraged both T-REx (Elsahar et al., 2018) and PARAREL to train BERT with the consistency loss and standard MLM loss. The fine-tuned BERT yielded impressive results when evaluated on the corresponding test dataset. However, it is worth noting that when applying the same fine-tuning approach to the SQuAD dataset (Gan and Ng, 2019), no significant impact was observed. Li et al. (2023c) initially employ the model as both a generator and a validator and queries the generator to acquire a response, following which the validator is consulted to assess the accuracy of the generated response. The paired responses from both the generator and the validator are filtered, retaining only those pairs that exhibit consistency, which is used to fine-tune the same model to enhance the likelihood of consistent pairs. They use their method to finetune Alpaca-30B with TriviaQA and demonstrate significant consistency enhancement in TriviaQA and natural questions.

Plug-in. Jang and Lukasiewicz (2023) first employ the intermediate training technique to retrain PLMs with word definition pairs in a dictionary to enhance the models' understanding of symbol meanings. They then introduce a training-efficient parameter integration method that merges the acquired parameters with those of other existing PLMs. They conduct experiments using RoBERTa as the backbone model and evaluate the effectiveness of their method in BECEL (Jang et al., 2022b).

Output Ensemble. Mitchell et al. (2022) propose a method to mitigate the inconsistency of LMs. They use a base model to create possible answers and a relation model to assess the logical relationships between these answers. The final answer is selected by considering both the base model's and the relation model's beliefs. The effectiveness of this approach on the language model is primarily demonstrated through experiments conducted on the BeliefBank QA dataset (Kassner et al., 2021). Instead, Zhao et al. (2023b) introduces a method to detect whether a question may cause inconsistency for LLMs. Specifically, they first use LLMs to rephrase the original question and obtain corresponding answers. They then cluster these answers and examine the divergence. The detection is determined based on the divergence level. They conduct comprehensive experiments using open-source and closed-source LLMs to validate the effectiveness of their proposed method on datasets including FaVI (Park et al., 2021) and ComQA (Abujabal et al., 2018).

A.3.2 Improving Factuality

Chuang et al. (2023) proposed a novel contrastive decoding approach named Dola. Specifically, the author initially developed a dynamic layer selection strategy, choosing the appropriate premature layers and mature layers. The next word's output probability is then determined by computing the difference in log probabilities of the premature layers and the mature layers. The author primarily conducted experiments on LLaMa in Truthful QA, Strategy QA, GSM8K, and FACTOR. The experiments provided substantial evidence that Dola consistently enhances the truthfulness of models and mitigates inconsistency issues. Li et al. (2023b) proposed a similar method named ITI. ITI first identifies a sparse set of attention heads that exhibit high linear probing accuracy for truthfulness, as measured by the TruthfulQA (Lin et al., 2022). During the inference phase, ITI shifts activations along the truth-correlated direction, obtained through knowledge probing. This intervention is repeated autoregressively for every token. The authors also verified their method using LLaMA-7B on TruthfulQA and found that ITI achieves a significant improvement in the factual knowledge accuracy of LLMs.