Knowledge Conflicts for LLMs: A Survey

Anonymous ARR submission

Abstract

This survey provides an in-depth analysis of knowledge conflicts for large language models (LLMs), highlighting the complex challenges they encounter when blending contextual and parametric knowledge. Our focus is on three categories of knowledge conflicts: context-memory, inter-context, and intra-memory conflict. These conflicts can significantly impact the trustworthiness and performance of LLMs, especially in real-world applications where noise and misinformation are common. By categorizing these conflicts, exploring the causes, examining the behaviors of LLMs under such conflicts, and reviewing available solutions, this survey aims to shed light on strategies for improving the robustness of LLMs, thereby serving as a valuable resource for advancing research in this evolving area.

1 Introduction

Large language models (LLMs; Brown et al. 2020; Touvron et al. 2023; OpenAI 2024) are renowned for encapsulating a vast repository of world knowledge (Petroni et al., 2019; Roberts et al., 2020), referred to as *parametric knowledge*. These models excel in various knowledge-intensive tasks. Meanwhile, LLMs continue to engage with external *contextual knowledge* after deployed (Pan et al., 2022), including user prompts (Liu et al., 2023a), documents from the Web (Shi et al., 2023c), or tools (Schick et al., 2023; Zhuang et al., 2023).

Integrating contextual knowledge into LLMs enables them to keep abreast of current events (Kasai et al., 2022) and generate more accurate responses (Shuster et al., 2021), yet it risks conflicting due to the rich knowledge sources. The discrepancies *among* the contexts and the model's parametric knowledge are referred to as *knowledge conflicts* (Chen et al., 2022; Xie et al., 2023). In this paper, we categorize **three** distinct types of knowledge conflicts, as shown in Figure 1. Contextual

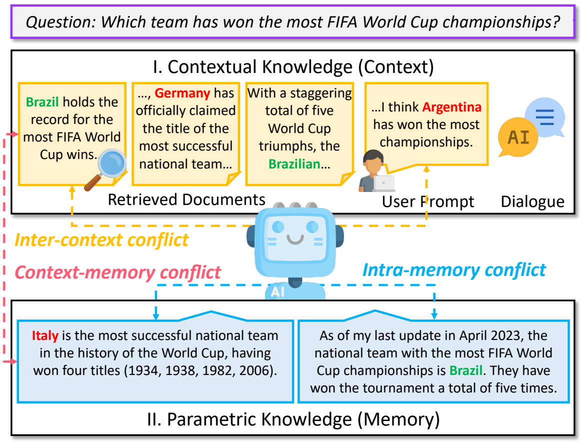

Figure 1: An LLM may encounter three types of knowledge conflicts, stemming from knowledge sources—either contextual (in yellow) or inherent to the LLM's parameters (in blue). When confronted with a user's question (in purple) entailing knowledge of complex conflicts, the LLM is required to resolve these discrepancies to deliver accurate responses.

knowledge (*context*, including user prompts, dialogue history, and retrieved documents) can conflict with the parametric knowledge (*memory*), where we term it as **context-memory conflict**. In the meantime, the context might be fraught with noise (Zhang and Choi, 2021) or even deliberately crafted misinformation (Du et al., 2022b). The conflict among contextual knowledge is dubbed as **inter-context conflict**. To reduce uncertainties in responses, the user may pose the question in various forms, resulting in the LLM's parametric knowledge in divergent responses. This variance may stem from the inconsistencies present in the pre-training data (Huang et al., 2023), which gives rise to what we call **intra-memory conflict**.

Knowledge conflicts attract attention with the advent of LLMs. 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 the context is factually wrong (Pan et al., 2023b). Given the im-

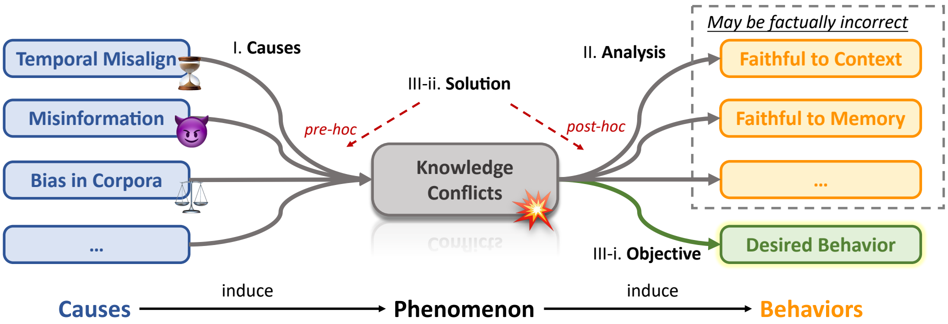

Figure 2: We view knowledge conflict not only as a standalone **phenomenon** but also as a nexus that connects various causal triggers (**causes**) with the **behaviors** of LLMs.

plications for the trustworthiness (Du et al., 2022b), real-time accuracy (Kasai et al., 2022), and robustness (Ying et al., 2023) of LLMs, it is imperative to delve deeper into understanding such conflicts (Xie et al., 2023; Wang et al., 2023e). Existing reviews (Zhang et al., 2023d; Wang et al., 2023a; Feng et al., 2023) either touch upon knowledge conflicts as a subtopic within a broader context and primarily focus on specific scenarios (Feng et al., 2023). To fill the gap, we aim to provide a comprehensive survey encompassing the categorization, cause and behavior analysis, and solutions for addressing various knowledge conflicts.

We conceptualize the *lifecycle of knowledge conflicts* as both a *cause* leading to various behaviors, and an *effect* emerges from the intricate nature of knowledge as in Figure 2. Our research underscores the significance of understanding the origins of these conflicts. Although existing analyses (Chen et al., 2022; Xie et al., 2023; Wang et al., 2023e) tend to construct such conflicts artificially, we posit that these analyses do not sufficiently address the interconnectedness of the issue. Going beyond, we provide a systematic review of mitigation strategies, which are employed to minimize the undesirable consequences of knowledge conflicts. Based on the timing relative to potential conflicts, such strategies are divided into *pre-hoc* and *post-hoc* strategies. The key distinction between them lies in whether adjustments are made *before* or *after* potential conflicts arise. We discuss three kinds of knowledge conflicts, detailing the causes, analysis of model behaviors, and available solutions according to their respective objectives. The taxonomy of knowledge conflicts is outlined in Figure 3.

2 Context-Memory Conflict

LLMs are characterized by fixed parametric knowledge, a result of the substantial pertaining process (Sharir et al., 2020; Hoffmann et al., 2022; Smith, 2023). This static parametric knowledge stands in stark contrast to the dynamic nature of external information, which evolves at a rapid pace (De Cao et al., 2021; Kasai et al., 2022).

2.1 Causes

Temporal Misalignment. It *naturally* arises in models trained on data collected in the past, as they may not accurately reflect contemporary realities (Luu et al., 2021; Lazaridou et al., 2021; Liska et al., 2022). Such misalignment can degrade the model's performance on various NLP tasks and relevancy over time (Luu et al., 2021; Zhang and Choi, 2021; Dhingra et al., 2022; Kasai et al., 2022; Cheang et al., 2023), as it may fail to capture new trends or shifts in language use. Furthermore, the issue of temporal misalignment is expected to intensify due to the pre-training paradigm and the escalating costs associated with scaling up models (Chowdhery et al., 2023; OpenAI, 2024).

Prior works tackle temporal misalignment by focusing on three lines of strategies: *Knowledge editing (KE)* aims to directly update the parametric knowledge (Sinitsin et al., 2020; Mitchell et al., 2021; Onoe et al., 2023). *Retrieval-augmented generation (RAG)* fetches relevant documents from external sources to supplement the model's knowledge without altering its parameters (Karpukhin et al., 2020; Guu et al., 2020; Lewis et al., 2020; Lazaridou et al., 2022; Vu et al., 2023). *Continual learning (CL)* updates the internal knowledge through continual training on updated data (Lazaridou et al., 2021; Jang et al., 2021, 2022). However, KE can bring in side effects such as knowledge inconsistency and may enhance the hallucination of LLMs (Li et al., 2023f; Pinter and Elhadad, 2023). RAG is inevitable to encounter conflicts since model parameters are not updated (Chen et al., 2021; Zhang and Choi, 2021). CL suffers from the issue of catastrophic forgetting and demands significant computational resources (De Lange et al., 2021; He et al., 2021; Wang et al., 2023d).

Misinformation Pollution. Adversaries can exploit this vulnerability by introducing misleading information into retrieved documents (Pan et al., 2023a,b; Weller et al., 2022) and user conversations (Xu et al., 2023). *Prompt injection* attack (Liu et al., 2023b; Greshake et al., 2023; Yi et al., 2023; Xu et al., 2024) is one such technique, where models may inadvertently spread misinformation if they use deceptive inputs (Pan et al., 2023b; Xu et al., 2023). Misinformation undermines the accuracy of automated fact-checking (Du et al., 2022b) and question-answering systems (Pan et al., 2023a,b). Recent studies 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). Recently, there has been growing apprehension regarding the potential generation of misinformation by LLMs (Ayoobi et al., 2023; Kidd and Birhane, 2023; Carlini et al., 2023; Zhou et al., 2023c; 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), which underscores the urgency of addressing the nuanced challenges LLMs pose within contextual misinformation.

Remarks. Temporal misalignment and misinformation pollution are two separate scenarios that give rise to context-memory conflicts. For the former, the up-to-date contextual information is considered accurate. *Conversely*, for the latter, the contextual information contains misinformation and is therefore considered incorrect.

2.2 Analysis of Model Behaviors

We summarize studies on how LLMs behave under context-memory conflicts within open-domain question answering (ODQA) and general setups.

ODQA. Early effort (Longpre et al., 2021) explores how QA models act when the provided contextual information contradicts the memory. An automated framework first 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. Longpre et al. (2021) reveal a tendency of models to over-rely on parametric knowledge. Chen et al. (2022) report differing observations, they note that models predominantly rely on contextual knowledge in their best-performing settings. This divergence can be attributed to two factors. Firstly, the entity substitution approach (Longpre et al., 2021) potentially reduces the semantic coherence of the perturbed context. Secondly, Chen et al. (2022) utilize multiple evidence rather than one (Longpre et al., 2021). Recently, Tan et al. (2024) examine how large LMs integrate context with generated memory. They observe that LLMs tend to prioritize parametric knowledge thanks to the greater similarity between generated contents and input, as well as the often incomplete nature of retrieved information.

General. LLMs exhibit a complex relationship with conflicting information. While highly receptive to convincing external evidence (Xie et al., 2023), they also demonstrate a strong confirmation bias (Nickerson, 1998), favoring information consistent with their memory. This leads to challenges in resolving such conflicts, as LLMs struggle to pinpoint conflicting segments and provide disentangled responses (Wang et al., 2023e). Research exploring LLMs' robustness under conflicts reveals a susceptibility to misleading prompts, particularly in commonsense knowledge (Ying et al., 2023). Furthermore, LLMs often deviate from their parametric knowledge when presented with direct conflicts or contextual changes (Qian et al., 2023). Studies investigating LLMs in interactive sessions highlight a tendency to favor logically structured knowledge, even when it is factual wrong (Xu et al., 2023). These findings underscore the need for further research into the interaction between parametric and contextual knowledge for LLMs.

Remarks. Researchers analyze LLMs' behavior under conflicting knowledge by creating artificial conflicts, initially through entity-level substitutions and later by using LLMs to generate semantically coherent conflicts. While no definitive rule exists for prioritizing contextual or parametric knowledge, LLMs tend to favor information that is semantically coherent over generic conflicting information.

2.3 Solutions

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

Faithful to Context. Several approaches have been proposed to achieve this goal. Fine-tuning approaches like Knowledge Aware (Li et al., 2022a) incorporate counterfactual and irrelevant contexts into training data to enhance controllability and robustness. Similarly, TrueTeacher (Gekhman et al., 2023) focus on improving factual consistency in summarization by annotating model-generated summaries with LLMs. Prompting strategies (Zhou et al., 2023d) utilize opinion-based prompts and counterfactual demonstrations to enhance LLMs' adherence to context without additional training. Decoding techniques like Context-aware Decoding (Shi et al., 2023a) amplify the difference in output probabilities with and without context, prioritizing relevant context over prior knowledge. Knowledge plug-in approaches, such as Continuously-updated QA (Lee et al., 2022a), use plug-and-play modules to store updated knowledge, solving knowledge conflicts without affecting the original model. Pre-training methods (Shi et al., 2023b) extend LLMs' ability to handle long and varied contexts across multiple documents, potentially resolving knowledge conflicts by synthesizing information from broader contexts. Finally, fact validity prediction approaches (Zhang and Choi, 2023) identify and discard outdated facts in LLMs, improving performance on tasks like ODQA by ensuring adherence to up-to-date contextual information.

Discriminating Misinformation. To combat misinformation, various defense strategies have been proposed. Pan et al. (2023b) advocates for misinformation detection and vigilant prompting, aiming to improve the model's faithfulness to factual information. Xu et al. (2023) employ a system prompt to encourage LLMs to be cautious about misinformation and verify their memorized knowledge before responding, further enhancing faithfulness. Weller et al. (2022) leverage the redundancy of information in large corpora to mitigate knowledge conflicts. Their approach involves query augmentation to retrieve diverse, less likely poisoned passages, then compares the consistency of predicted answers across retrieved contexts. This strategy ensures faithfulness by cross-verifying answers from multiple sources. Hong et al. (2023) fine-tune a smaller LM as a discriminator and integrate prompting techniques to enable the model to distinguish between reliable and unreliable information.

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. (2023e) introduce a method to improve 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.

Improving Factuality. Zhang et al. (2023e) 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. Jin et al. (2024a) introduces a contrastive-decoding-based algorithm to maximize the difference between various logits under knowledge conflicts and calibrates the model's confidence in the truthful answer.

Remarks. Current mitigation approaches for knowledge conflicts are ineffective because they fail to differentiate between the two underlying causes. Blindly prioritizing either faithfulness to context or knowledge is undesirable. Researchers advocate for LLMs that empower users to make informed decisions by providing distinct answers based on both parametric and contextual information (Wang et al., 2023e; Floridi, 2023).

3 Inter-Context Conflict

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

3.1 Causes

Misinformation. Similar to context-memory conflict, this type of conflict can also affected by misinformation and will not be discussed repeatedly.

Outdated Information. It is also important to recognize that facts can evolve. Retrieved documents may contain updated and outdated information from the network simultaneously, leading to conflicts between these documents (Chen et al., 2021; Liska et al., 2022; Kasai et al., 2022).

3.2 Analysis of Model Behaviors

Performance Impact. Previous research has shown that LMs can be significantly influenced by misinformation or outdated information within a specific context (Zhang and Choi, 2021; Du et al., 2022b). Pan et al. (2023a) demonstrated that LLMs are susceptible to misinformation attacks, even when the fake articles are generated by models. Chen et al. (2022) investigated how LLMs handle contradictory contexts and found that inconsistencies across knowledge sources have a minimal effect on their confidence levels. These models tend to favor context directly related to the query and context that aligns with their parametric knowledge. Xie et al. (2023) confirmed these findings, showing that LLMs exhibit a bias towards evidence that aligns with their parametric memory and a predisposition towards emphasizing information related to popular entities and answers corroborated by a larger volume of documents. Furthermore, they found that LLMs are sensitive to the order in which data is introduced. Jin et al. (2024a) discovered that LLMs struggle with reasoning as the number of conflicting hops increases.

Detection Ability. Several studies highlight the challenges faced by LMs in identifying contradictions. Zheng et al. (2022) demonstrate that LMs struggle to detect contradictory statements within Chinese conversations. Li et al. (2023a) analyze the performance of LLMs in identifying contradictory documents across various sources, including news (Hermann et al., 2015), stories (Kočiský et al., 2018), and Wikipedia (Merity et al., 2017), finding that the average detection accuracy is low. They also observe that LLMs perform poorly when dealing with contradictions involving subjective emotions or perspectives. Wan et al. (2024) investigate the text features influencing LLMs' assessment of document credibility in the presence of conflicting information, discovering that models prioritize relevance over stylistic features. Jin et al. (2024a) further highlight the difficulty LLMs encounter in distinguishing truthful information from misinformation, showing a tendency to favor evidence that appears most frequently within the context.

Remarks. Exploring responses to contextual nuances is essential, as variations in training data lead to differences in behavior. Despite some similarities, LLMs' methods of identifying misinformation differ significantly from those of humans.

3.3 Solutions

Eliminating Conflict. Several approaches have been proposed to address the challenge of eliminating conflict in text. Specialized models, such as the Pairwise Contradiction Neural Network (Hsu et al., 2021), utilize fine-tuned Sentence-BERT embeddings to determine contradiction probabilities. Pielka et al. (2022) emphasize the importance of integrating linguistic knowledge into the learning process to improve contradiction detection, as models like XLM-RoBERTa struggle with syntactic and semantic features. Wu et al. (2022) propose incorporating topological text representations into language models to enhance contradiction detection, evaluating their approach on the MultiNLI dataset (Williams et al., 2018). General models, such as Chern et al. (2023)'s fact-checking framework, integrate LLMs with various tools to detect factual errors. Leite et al. (2023) leverage LLMs to generate weak labels associated with credibility signals for input text, aggregating these labels through weak supervision techniques to predict veracity.

Improving Robustness. To enhance robustness, Hong et al. (2023) propose a fine-tuning method that trains a discriminator and decoder simultaneously using a shared encoder, alongside strategies involving prompting GPT-3 to identify perturbed documents and integrating the discriminator's output into prompts. Weller et al. (2022) explore query augmentation by prompting GPT-3 to generate new questions based on the original query, evaluating answer confidence through passage retrieval, and deciding whether to rely on the original prediction or aggregate predictions from high-confidence augmented questions. While both approaches aim for robustness, Hong et al. (2023)'s fine-tuning method demonstrates the most promising results.

Remarks. Strategies for addressing inter-context conflicts primarily rely on model knowledge or leverage external knowledge such as retrieved documents. Moreover, augmenting LLM capabilities with external tools has emerged as a novel paradigm. Exploring the use of external tools to support LLMs in resolving inter-context conflicts is a promising approach. In addition, devising a unified and efficient approach to handle various conflict types remains a formidable challenge.

4 Intra-Memory Conflict

Consistent LLM outputs for identical inputs are essential. However, intra-memory conflicts, where LLMs generate differing responses to similar inputs, undermine their reliability and utility by introducing undesirable uncertainty.

4.1 Causes

The following three factors respectively pertain to training, inference, and knowledge refinement.

Bias in Training Corpora. While LLMs primarily acquire knowledge during pre-training (Zhou et al., 2023a; Kaddour et al., 2023; Naveed et al., 2023; Akyürek et al., 2022; Singhal et al., 2022), the vast and often unreliable nature of internet-sourced training data (Bender et al., 2021; Weidinger et al., 2021) can lead to the memorization and amplification of inaccuracies (Lin et al., 2022; Elazar et al., 2022; Lam et al., 2022; Grosse et al., 2023). This results in LLMs potentially harboring conflicting knowledge within their parameters. Furthermore, LLMs tend to encode superficial associations rather than true comprehension of training data (Li et al., 2022b; Kang and Choi, 2023; Zhao et al., 2023a; Kandpal et al., 2023), leading to predetermined responses based on spurious correlations and potentially divergent answers for semantically equivalent but syntactically distinct prompts.

Decoding Strategy. LLMs generate text by sampling from a probability distribution over potential next tokens. Stochastic sampling methods like top-k and top-p sampling are commonly used for decoding, introducing randomness in the generated content (Jawahar et al., 2020; Massarelli et al., 2020; Fan et al., 2018; Holtzman et al., 2020). However, this randomness can cause intra-memory conflicts, where the model produces different outputs for the same input due to the left-to-right generation pattern and the influence of sampled tokens on subsequent generations (Lee et al., 2022b; Huang et al., 2023; Dziri et al., 2021).

Knowledge Editing. With the exponential increase of model parameters, fine-tuning LLMs become increasingly resource-intensive. In response to this, researchers explore knowledge editing techniques to efficiently modify a small scope of the knowledge in LLMs (Meng et al., 2022; Zhong et al., 2023). Ensuring the consistency of such modification poses a significant challenge. Due to the potential limitations inherent in the editing method, the modified knowledge cannot be generalized effectively. This can result in LLMs producing inconsistent responses when dealing with the same piece of knowledge in varying situations (Li et al., 2023f; Yao et al., 2023).

Remarks. Intra-memory conflicts in LLMs arise from three main causes at different stages. Training corpus bias is the primary catalyst, causing inconsistencies in the model's knowledge. The randomness of the decoding process during inference exacerbates these inconsistencies. Additionally, knowledge editing can inadvertently introduce conflicting information.

4.2 Analysis of Model Behaviors

Self-Inconsistency. LLMs exhibit significant self-inconsistency, as evidenced by multiple studies. Elazar et al. (2021) found that BERT, RoBERTa, and ALBERT struggle with knowledge consistency, achieving accuracy rates barely exceeding 50-60%. Hase et al. (2023), using a more diverse dataset, confirmed these findings, highlighting the inconsistency of RoBERTa-base and BART-base in paraphrase contexts. Zhao et al. (2023b) revealed that even GPT-4 displays a 13% inconsistency rate in Commonsense Question-Answering tasks, particularly when dealing with uncommon knowledge. Dong et al. (2023) further demonstrated that various open-source LLMs exhibit strong inconsistencies. Li et al. (2023d) identified another aspect of inconsistency, where LLMs may initially answer a question but subsequently deny the answer when asked for confirmation. Li et al. (2022b) attributed this inconsistency in encoder-based models to their reliance on positionally close and highly co-occurring words, leading to the generation of misinformation. Kang and Choi (2023) further explained this phenomenon as a co-occurrence bias, where LLMs prioritize frequently co-occurring words over correct answers, particularly when recalling facts with rarely co-occurring subject-object pairs in the pre-training dataset, even after fine-tuning.

Latent Representation of Knowledge. Contemporary LLMs, built on multi-layer transformer architectures, exhibit a complex inter-memory conflict with distinct knowledge representations scattered across layers. Research suggests that LLMs store low-level information at shallower layers and semantic information at deeper layers (Tenney et al., 2019; Rogers et al., 2020; Wang et al., 2019; Jawahar et al., 2019; Cui et al., 2020). Chuang et al. (2023) demonstrate that factual knowledge is concentrated within specific transformer layers, leading to inconsistent knowledge across layers. Furthermore, Li et al. (2023c) highlight a discrepancy between knowledge storage and generation accuracy. Their experiments reveal a 40% gap between the accuracy of a knowledge probe and the generation accuracy, suggesting that while the correct knowledge is present within the parameters, it may not be effectively expressed during generation.

Cross-lingual Inconsistency. While true knowledge should be universally accessible regardless of language variation (Ohmer et al., 2023), LLMs exhibit cross-lingual inconsistencies (Ji et al., 2023; Xue et al., 2024). This inconsistency arises from LLMs storing knowledge related to different languages separately within their parameters (Wang et al., 2023c). Qi et al. (2023) propose RankC, a metric for evaluating cross-lingual consistency of factual knowledge, and reveals a strong language dependence in LLMs, with no improvement in consistency observed even with larger models.

Remarks. The phenomenon of inter-memory conflict in LLMs predominantly manifests through inconsistent responses to semantically identical queries. This inconsistency is primarily attributed to the suboptimal quality of datasets utilized during the pre-training phase. Addressing this challenge necessitates the development of efficient and cost-effective solutions, which remains a significant hurdle. Additionally, LLMs are characterized by the presence of multiple knowledge circuits, which significantly influence their response mechanisms to specific inquiries. The exploration and detailed examination of these knowledge circuits within LLMs represent a promising avenue for future research.

4.3 Solutions

Improving Consistency. Several approaches have been proposed to address the inconsistency issue in language models. Fine-tuning methods, such as those explored by Elazar et al. (2021) and Li et al. (2023d), aim to improve consistency by introducing loss functions that penalize inconsistent outputs or by selectively retaining only consistent response pairs for training. Jang and Lukasiewicz (2023) propose a plug-in method that leverages intermediate training with word-definition pairs to enhance the model's understanding of symbolic meanings, thereby mitigating inconsistency. Output ensemble approaches, such as those presented by Mitchell et al. (2022) and Zhao et al. (2023b), utilize multiple models to evaluate the consistency of generated outputs. Mitchell et al. (2022) employ a base model for generating potential answers and a relation model for assessing their logical coherence, while Zhao et al. (2023b) leverage LLMs to rephrase questions and analyze the divergence of corresponding answers to detect potential inconsistency. These diverse approaches highlight the ongoing efforts to enhance the consistency and reliability of language models.

Improving Factuality. Chuang et al. (2023) and Li et al. (2023c) propose methods that leverage the inconsistency of knowledge across different layers. DoLa (Chuang et al., 2023) utilizes a dynamic layer selection strategy, contrasting premature and mature layers to determine the next word's probability. ITI (Li et al., 2023c), on the other hand, identifies truth-correlated attention heads based on TruthfulQA (Lin et al., 2022) and shifts activations along this direction during inference, repeating this process autoregressively for each token. Both approaches aim to mitigate factual errors by effectively utilizing the diverse knowledge representations within the model's layers.

Remarks. The resolution of inter-memory conflict in LLMs typically entails three phases: training, generation, and post-hoc processing. The training phase method mainly focuses on mitigating internal inconsistencies among model parameters. Conversely, the generation and post-hoc phases primarily involve algorithmic interventions aimed at alleviating occurrences of inconsistent model behavior. Nevertheless, the challenge persists in addressing the inconsistency of parameter knowledge without detrimentally impacting the overall performance of LLMs.

5 Challenges and Future Directions

Knowledge Conflicts in the Wild. While current research on knowledge conflicts primarily focuses on artificially generated misinformation, real-world conflicts often arise in retrieval-augmented LLMs due to conflicting information retrieved from the web. Existing analyses lack the realism of such scenarios, potentially limiting the applicability of their findings (Xie et al., 2023; Wang et al., 2023e). Recent work has begun to address this gap by curating conflicting documents based on actual Google search results for open-ended questions (Wan et al., 2024). Future research should prioritize evaluating LLMs in these real-world scenarios to better understand their capabilities and limitations.

Solution at a Finer Resolution. Resolving knowledge conflicts presents a complex challenge, lacking a universal solution. Conflicting information can stem from misinformation, outdated facts, or partially correct data (Uscinski and Butler, 2013; Guo et al., 2022). Existing approaches often rely on simple prior assumptions (Shi et al., 2023b). A more nuanced approach is desired, considering the query's nature, the type of conflict, and user expectations (Floridi, 2023), *e.g.*, subjective or debatable questions inherently lead to conflicts due to multiple valid answers (Bjerva et al., 2020; Wan et al., 2024). Future solutions should acknowledge the diverse causes, manifestations, and potential user expectations, requiring collaboration between NLP and social science researchers for comprehensive investigation and effective solutions.

Evaluation on Downstream Tasks. While research on knowledge conflicts primarily focuses on evaluating their performance on QA datasets, the broader implications of these conflicts remain underexplored. Their impact on downstream tasks, particularly those demanding high accuracy and consistency, such as legal document analysis (Shui et al., 2023; Martin et al., 2024), medical diagnosis (Zhou et al., 2023b; Thirunavukarasu et al., 2023), financial analysis (Zhang et al., 2023a; Li et al., 2023e), and educational tools (Caines et al., 2023; Milano et al., 2023), is crucial. Unresolved knowledge conflicts could severely hinder the utility of these models in such applications.

Interplay among the Conflicts. Current research primarily focuses on individual conflict types or a combined study of inter-context and context-memory conflicts. However, the interplay between intra-memory conflict and other types of conflicts remains unexplored. Notably, several studies have proposed the existence of knowledge circuits in LLMs (Chughtai et al., 2024; Huang et al., 2023), which are closely related to intra-memory conflict. Understanding this interaction is crucial for comprehending the relationship between internal knowledge inconsistency and model behavior in response to context. Furthermore, exploring the synergistic effects of various conflict types could reveal underlying mechanisms of knowledge representation and processing in LLMs is vital.

Explainability. While research has focused on analyzing LLMs' outputs when faced with knowledge conflicts, the internal mechanisms driving these decisions remain underexplored. Studies examining model confidence through logits (Xu et al., 2023; Jin et al., 2024a; Wang et al., 2024) offer some insights, but a deeper understanding of how specific attention heads or neuron activations contribute to conflict resolution is needed. Jin et al. (2024b) made progress by investigating the interpretability of LLMs through information flow analysis, identifying memory and context heads with opposing effects in later layers. However, further microscopic examinations are required to fully comprehend how LLMs navigate conflicting information.

Multilinguality. Current research has primarily focused on English. Future research should expand to address conflicts in non-English texts, leveraging multilingual LLMs like GPT-4 (OpenAI, 2024) and GLM (Zeng et al., 2022) to account for language-specific characteristics. Additionally, inter-context conflict, involving documents in different languages, requires solutions like translation systems (Dementieva and Panchenko, 2021), leveraging high-resource language evidence for low-resource languages (Xue et al., 2024), or employing knowledge distillation techniques.

Multimodality. While current research mainly focuses on text modality, potential conflicts arises as LLMs evolve to process information across various formats, including text, images (Alayrac et al., 2022; Li et al., 2023b), video (Ju et al., 2022; Zhang et al., 2023b), and audio (Borsos et al., 2023; Wu et al., 2023). For example, an audio clip might contradict an accompanying document. Future research should focus on enhancing models' ability to navigate these complex multimodal dynamics, developing targeted datasets for training and evaluation, and exploring user perception of multimodal conflicts to improve LLMs.

6 Conclusion

This paper delves into the multifaceted issue of knowledge conflicts, analyzing the categorization, causes, behavior, and mitigation. We demonstrate that the type of conflict significantly influences a model's behavior and that these conflicts exhibit complex interplays. Existing solutions, often focused on artificial scenarios and relying on priors, lack the granularity and breadth needed to address the increasing complexity of knowledge conflicts in real-world applications. As retrieval-augmented LLMs become more prevalent, comprehensive research on knowledge conflicts is crucial.

Limitations

Considering the rapid expansion of research in the field of knowledge conflict and the abundance of scholarly literature, it is possible that we might have missed some of the most recent or less relevant findings. Nevertheless, we have ensured the inclusion of all essential materials in our survey.

Ethics Statement

We mainly searched for papers published after 2021 using key terms including "knowledge conflict", "knowledge inconsistency", "knowledge gap", *inter alia*, on Google Scholar and the ACL Anthology. After initially identifying these papers, the authors classified them through reading and continued to track related but overlooked papers using their citations. We also used Google Scholar to follow up on the latest papers citing these to avoid omissions.

For the quantitative analysis and comparison section (§ F), we did not conduct computational experiments but simply organized the result reported in other literature as is.

References

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, Online. Association for Computational Linguistics.

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

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

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.

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.

Johannes Bjerva, Nikita Bhutani, Behzad Golshan, Wang-Chiew Tan, and Isabelle Augenstein. 2020. SubjQA: A Dataset for Subjectivity and Review Comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5480–5494, Online. Association for Computational Linguistics.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. 2023. Audiolm: a language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Andrew Caines, Luca Benedetto, Shiva Taslimipoor, Christopher Davis, Yuan Gao, Oeistein Andersen, Zheng Yuan, Mark Elliott, Russell Moore, Christopher Bryant, et al. 2023. On the application of large language models for language teaching and assessment technology. *ArXiv preprint*, abs/2307.08393.

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*, abs/2302.10149.

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 Work-*

shop on Instruction Tuning and Instruction Following.

Canyu Chen and Kai Shu. 2023b. Combating misinformation in the age of llms: Opportunities and challenges. *ArXiv preprint*, abs/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*, abs/2210.13701.

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

Wenhu Chen, Xinyi Wang, and William Yang Wang. 2021. A dataset for answering time-sensitive questions. *ArXiv preprint*, abs/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*, abs/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*, abs/2309.03883.

Bilal Chughtai, Alan Cooney, and Neel Nanda. 2024. Summing up the facts: Additive mechanisms behind factual recall in llms. *ArXiv preprint*, abs/2402.07321.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Leyang Cui, Sijie Cheng, Yu Wu, and Yue Zhang. 2020. Does bert solve commonsense task via commonsense knowledge. *ArXiv preprint*, abs/2008.03945.

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.

Daryna Dementieva and Alexander Panchenko. 2021. Cross-lingual evidence improves monolingual fake news detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 310–320, Online. Association for Computational Linguistics.

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*, abs/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*, abs/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*, abs/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)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

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*, abs/2311.05876.

Luciano Floridi. 2023. Ai as agency without intelligence: on chatgpt, large language models, and other generative models. *Philosophy & Technology*, 36(1):15.

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

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*, abs/2308.03296.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, 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, Online. Association for Computational Linguistics.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1693–1701.

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*, abs/2203.15556.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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*, abs/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*, abs/2311.05232.

Joel Jang, Seonghyeon Ye, Changho Lee, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, and Minjoon Seo. 2022. 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 Erik Jang and Thomas Lukasiewicz. 2023. Improving language models meaning understanding and consistency by learning conceptual roles from dictionary. *ArXiv preprint*, abs/2310.15541.

Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks Lakshmanan, V.S. 2020. Automatic detection of machine generated text: A critical survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2296–2309, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. 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*, abs/2309.15847.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Qiuxia Li, and Jun Zhao. 2024a. Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. *ArXiv preprint*, abs/2402.14409.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2024b. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. *ArXiv preprint*, abs/2402.18154.

Mandar Joshi, Eunsol Choi, Daniel 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, Vancouver, Canada. Association for Computational Linguistics.

Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. 2022. Prompting visual-language models for efficient video understanding. In *European Conference on Computer Vision*, pages 105–124. Springer.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models. *ArXiv preprint*, abs/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*, abs/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, Online. Association for Computational Linguistics.

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*, abs/2207.13332.

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*, abs/2205.12221.

Tomáš Kočiský, 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.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

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*, abs/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*, abs/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*, abs/2309.07601.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

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*, abs/2211.05110.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

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

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

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2022c. Contrastive decoding: Open-ended text generation as optimization. *ArXiv preprint*, abs/2210.15097.

Xiang Lisa Li, Vaishnavi Shrivastava, Siyan Li, Tatsunori Hashimoto, and Percy Liang. 2023d. Benchmarking and improving generator-validator consistency of language models. *ArXiv preprint*, abs/2310.01846.

Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2023e. Large language models in finance: A survey. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pages 374–382.

Zhoubo Li, Ningyu Zhang, Yunzhi Yao, Mengru Wang, Xi Chen, and Huajun Chen. 2023f. Unveiling the pitfalls of knowledge editing for large language models. *ArXiv preprint*, abs/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*, abs/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*, abs/2111.07408.

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*, abs/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*, abs/2302.12297.

Lauren Martin, Nick Whitehouse, Stephanie Yiu, Lizzie Catterson, and Rivindu Perera. 2024. Better call gpt, comparing large language models against lawyers. *ArXiv preprint*, abs/2401.16212.

Luca Massarelli, Fabio Petroni, Aleksandra Piktus, Myle Ott, Tim Rocktäschel, Vassilis Plachouras, Fabrizio Silvestri, and Sebastian Riedel. 2020. How decoding strategies affect the verifiability of generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 223–235, Online. Association for Computational Linguistics.

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. 2017. Pointer sentinel mixture models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Silvia Milano, Joshua A McGrane, and Sabina Leonelli. 2023. Large language models challenge the future of higher education. *Nature Machine Intelligence*, 5(4):333–334.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2021. Fast model editing at scale. *ArXiv preprint*, abs/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*, abs/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*, abs/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*, abs/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*, abs/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*, abs/2305.01651.

OpenAI. 2024. Gpt-4 technical report.

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*, abs/2305.13661.

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*, abs/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, Hong Kong, China. Association for Computational Linguistics.

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*, abs/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*, abs/2309.08594.

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)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. 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*, abs/2302.04761.

Or Sharir, Barak Peleg, and Yoav Shoham. 2020. The cost of training nlp models: A concise overview. *ArXiv preprint*, abs/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*, abs/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*, abs/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*, abs/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*, abs/2301.12652.

Ruihao Shui, Yixin Cao, Xiang Wang, and Tat-Seng Chua. 2023. A comprehensive evaluation of large language models on legal judgment prediction. *ArXiv preprint*, abs/2310.11761.

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*, abs/2212.13138.

Anton Sinitsin, Vsevolod Plokhotnyuk, Dmitriy Pyrkin, Sergei Popov, and Artem Babenko. 2020. Editable neural networks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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

Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. 2023. Ai model gpt-3 (dis) informs us better than humans. *ArXiv preprint*, abs/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*, abs/2401.11911.

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

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

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*, abs/2307.09288.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multihop 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*, abs/2305.04388.

Joseph E Uscinski and Ryden W Butler. 2013. The epistemology of fact checking. *Critical Review*, 25(2):162–180.

Tyler Vergho, Jean-Francois Godbout, Reihaneh Rabbany, and Kellin Pelrine. 2024. Comparing gpt-4 and open-source language models in misinformation mitigation. *ArXiv preprint*, abs/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*, abs/2310.03214.

Alexander Wan, Eric Wallace, and Dan Klein. 2024. What evidence do language models find convincing? *ArXiv preprint*, abs/2402.11782.

Cunxiang Wang, Shuailong Liang, Yue Zhang, Xiaonan Li, and Tian Gao. 2019. Does it make sense? and why? a pilot study for sense making and explanation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4020–4026, Florence, Italy. Association for Computational Linguistics.

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.

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

Hongru Wang, Boyang Xue, Baohang Zhou, Tianhua Zhang, Cunxiang Wang, Guanhua Chen, Huimin Wang, and Kam fai Wong. 2024. Self-dc: When to retrieve and when to generate? self divide-and-conquer for compositional unknown questions.

Jiaan Wang, Yunlong Liang, Zengkui Sun, Yuxuan Cao, and Jiarong Xu. 2023c. Cross-lingual knowledge editing in large language models. *ArXiv preprint*, abs/2309.08952.

Liyuan Wang, Xingxing Zhang, Qian Li, Mingtian Zhang, Hang Su, Jun Zhu, and Yi Zhong. 2023d. 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. 2023e. Resolving knowledge conflicts in large language models. *ArXiv preprint*, abs/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*, abs/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*, abs/2112.04359.

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

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

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*, abs/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*, abs/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*.

Rongwu Xu, Zehan Qi, and Wei Xu. 2024. Preemptive answer" attacks" on chain-of-thought reasoning. *arXiv preprint arXiv:2405.20902*.

Boyang Xue, Hongru Wang, Weichao Wang, Rui Wang, Sheng Wang, Zeming Liu, and Kam-Fai Wong. 2024. A comprehensive study of multilingual confidence estimation on large language models.

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*, abs/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*, abs/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*, abs/2309.17415.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations*.

Boyu Zhang, Hongyang Yang, Tianyu Zhou, Muhammad Ali Babar, and Xiao-Yang Liu. 2023a. Enhancing financial sentiment analysis via retrieval augmented large language models. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pages 349–356.

Hang Zhang, Xin Li, and Lidong Bing. 2023b. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 543–553.

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

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

Michael JQ Zhang and Eunsol Choi. 2023. Mitigating temporal misalignment by discarding outdated facts. *ArXiv preprint*, abs/2305.14824.

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. 2023d. 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. 2023e. Merging generated and retrieved knowledge for open-domain qa. *ArXiv preprint*, abs/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*, abs/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*, abs/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*, abs/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*, abs/2305.11206.

Hongjian Zhou, Boyang Gu, Xinyu Zou, Yiru Li, Sam S Chen, Peilin Zhou, Junling Liu, Yining Hua, Chengfeng Mao, Xian Wu, et al. 2023b. A survey of large language models in medicine: Progress, application, and challenge. *ArXiv preprint*, abs/2311.05112.

Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023c. 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. 2023d. Context-faithful prompting for large language models. *ArXiv preprint*, abs/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*, abs/2306.13304.

A Taxonomy of Knowledge Conflicts

Figure 3 outlines the taxonomy we used in organize this survey. To start with, we classify knowledge conflicts into three categories based on the sources: context-memory conflict (§ 2), inter-context conflict (§ 3), and intra-memory conflict (§ 4). Within each type of conflict, we sequentially present its causes, analysis of LLMs' behaviors, and possible mitigation solutions. Each specific issue is further categorized according to its internal characteristics (*e.g.*, solutions are categorized based on the characteristics of the strategies engaged).

B Datasets of Knowledge Conflicts

We list notable datasets employed in investigating the three types of knowledge conflict in Table 1. It is worth noting that for all context-memory datasets, extra attention should be paid to their applicability. This is because these datasets always need to be based on model-specific memories as a baseline when constructing conflicting knowledge. Obviously, this parameterized knowledge varies from model to model, greatly reducing the reusability of these datasets. Furthermore, the value of these datasets is further diminished by the existence of model variants from different *knowledge cutoff date* (*e.g.*, OpenAI's GPT-4 family of models). The parameterized knowledge varies from variant to variant due to different cutoff date.

C Detailed Solutions for Context-Memory Conflict

C.1 Faithful to Context

Fine-tuning. Li et al. (2022a) argue 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 FineTuning (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

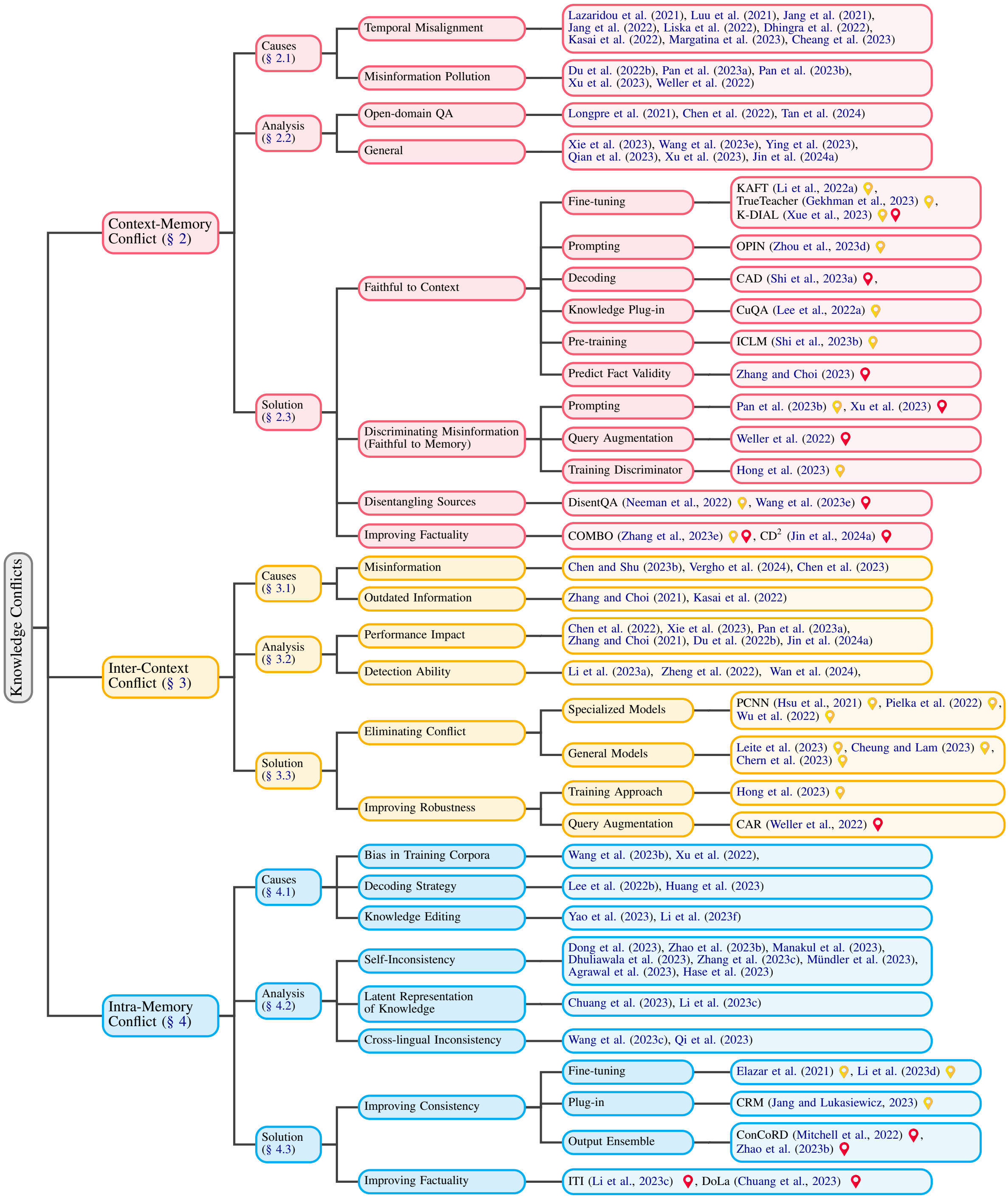

Figure 3: Taxonomy of knowledge conflicts. We mainly list works in the era of large language models. (yellow marker) denotes pre-hoc solution and (red marker) denotes post-hoc solution.

dialogue systems via direct knowledge enhancement and reinforcement learning for factual consistency (RLFC) for aligning responses accurately with provided factual knowledge.

Prompting. Zhou et al. (2023d) 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) introduce Context-aware Decoding (CAD) to reduce hallucinations by amplifying the difference in output probabilities with and without context, which is similar to the concept of contrastive decoding (Li et al., 2022c). CAD enhances faithfulness in LLMs by effectively

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

1. Approach refers to how the conflicts are crafted, including entity-level substitution (Sub), generative approaches employing an LLM (Gen), and 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, one should select a subset of the dataset that aligns with the tested model's knowledge when using CM datasets.

Table 1: Datasets on evaluating a large language model's behavior when encountering knowledge conflicts. CM: context-memory conflict, IC: inter-context conflict, IM: intra-memory conflict.

prioritizing relevant context over the model's prior knowledge, especially in tasks with conflicting information.

Knowledge Plug-in. Lee et al. (2022a) propose 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.

C.2 Discriminating Misinformation (Faithful to Memory)

Prompting. To address misinformation pollution, Pan et al. (2023b) propose 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) utilize 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) leverage 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 the cross-verification of answers in multiple sources.

Training Discriminator. Hong et al. (2023) fine-tune a smaller LM as a 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.

C.3 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. (2023e) introduce a method to improve 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.

C.4 Improving Factuality

Zhang et al. (2023e) 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. Jin et al. (2024a) introduce a contrastive-decoding-based algorithm, namely CD^2, which maximizes the difference between various logits under knowledge conflicts and calibrates the model's confidence in the truthful answer.

D Detailed Solutions for Inter-Context Conflict

D.1 Eliminating Conflict

Specialized Models. Hsu et al. (2021) develop a model named Pairwise Contradiction Neural Network (PCNN), leveraging fine-tuned Sentence-BERT embeddings to calculate contradiction probabilities of articles. Pielka et al. (2022) suggest incorporating linguistic knowledge into the learning process based on the discovery that XLM-RoBERTa struggles to effectively grasp the syntactic and semantic features that are vital for accurate contradiction detection. Wu et al. (2022) propose an innovative approach that integrates topological representations of text into language models to enhance the contradiction detection ability and evaluated their methods on the MultiNLI dataset (Williams et al., 2018).

General Models. Chern et al. (2023) propose a fact-checking framework that integrates LLMs with various tools, including Google Search, Google Scholar, code interpreters, and Python, for detecting factual errors in texts. Leite et al. (2023) employ LLMs to generate weak labels associated with predefined credibility signals for the input text and aggregate these labels through weak supervision techniques to make predictions regarding the veracity of the input.

D.2 Improving Robustness

Training Approach. Hong et al. (2023) present a novel fine-tuning method that involves training a discriminator and a decoder simultaneously using a shared encoder. Additionally, the authors introduce two other strategies 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. Their experimental results indicate that the fine-tuning method yields the most promising results.

Query Augmentation. Weller et al. (2022) explore a query augmentation technique that prompts GPT-3 to formulate new questions derived from the original inquiry. They then assess the confidence for each question's answer by referencing the corresponding passages retrieved. Based on the confidence, they decide whether to rely on the original question's prediction or aggregate predictions from the augmented questions with high confidence scores.

E Detailed Solutions for Intra-Memory Conflict

E.1 Improving Consistency

Fine-tuning. Elazar et al. (2021) propose a consistency loss function and train the language model with the combination of the consistency loss and standard MLM loss. Li et al. (2023d) utilize one language model in dual capacities: as a generator to produce responses and as a validator to evaluate the accuracy of these responses. The process involves querying the generator for a response, which is subsequently assessed by the validator for accuracy. Only those pairs of responses deemed consistent are retained. This subset of consistent pairs is then used to fine-tune the model, aiming to increase the generation likelihood of consistent response pairs.

Plug-in. Jang and Lukasiewicz (2023) leverage the technique of intermediate training, utilizing word-definition pairs from dictionaries to retrain language models and improve their comprehension of symbolic meanings. Subsequently, they propose an efficient parameter integration approach, which amalgamates these enhanced parameters with those of existing language models. This method aims to rectify the models' inconsistent behavior by bolstering their capacity to understand meanings.

Output Ensemble. Mitchell et al. (2022) propose a method to mitigate the inconsistency of language models by leveraging a two-model architecture, involving the utilization of a base model responsible for generating a set of potential answers, followed by a relation model that evaluates the logical coherence among these answers. The final answer is selected by considering both the base model's and the relation model's beliefs. Zhao et al. (2023b) introduce 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.

E.2 Improving Factuality

Chuang et al. (2023) propose a novel contrastive decoding approach named DoLa. Specifically, the authors develop 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. Li et al. (2023c) devise a similar method named ITI. They first identify a sparse set of attention heads that exhibit high linear probing accuracy for truthfulness, as measured by TruthfulQA (Lin et al., 2022). During the inference phase, ITI shifts activations along the truth-correlated direction, which is obtained through knowledge probing. This intervention is repeated autoregressively for every token during completion. Both DoLa and ITI address the inconsistency of knowledge across the model's different layers to reduce factual errors.

F Quantitative Analysis and Comparison

In the context of a survey paper, while it is beneficial to include quantitative results and analyses concerning the impact of knowledge conflicts across various types of conflicts and the performance comparison of different mitigation strategies, it is not a strict requirement. We acknowledge the *complexity and impracticality* involved in conducting such quantitative experiments, particularly due to the use of disparate datasets in behavioral analyses, as well as the variance in the inherent knowledge of LLMs across different knowledge cut-off snapshots, as detailed in § B.

Moreover, establishing a "fair" comparison within the mitigation strategies segment poses its own set of challenges, given the diversity in objectives influenced by various assumed priors, such as the perceived accuracy of context or inherent knowledge, as discussed in the main text. Despite these intricacies, we opt to present quantitative results by compiling existing evaluations from a range of papers. *It is imperative, however, to* ***approach this analysis with caution****, recognizing that original authors may have employed different datasets, LLMs variants, or even pursued contrasting objectives.*

F.1 Quantitative Results on the Impact of Knowledge Conflicts

The comparison of quantitative results on the impact of the three types of knowledge conflicts is shown in Table 2. We pick the results of representative behavior analysis literature for comparison.

F.2 Quantitative Results on the Effectiveness of Mitigation Strategies

The effectiveness of various mitigation strategies is quantitatively compared in Table 3. It is important to note that our analysis is limited to works addressing *three predominant types of mitigating objectives* within the context of memory conflicts. This selection is deliberate, as other types of mitigating objectives in different conflict categories do not yet have a substantial body of work that would allow for a meaningful cross-method comparison.

Reference	Model	Dataset	Quantitative Results
		Context-memory conflict	
Pan et al. (2023b)	ChatGPT	NQ-1500 and CovidNews	Misinformation in the context can lead to a significant degradation (up to 87%) in the performance.
Xie et al. (2023)	ChatGPT, GPT-4, PaLM2, Qwen, Llama2, and Vicuna	POPQA and STRATEGYQA	For entity substitution-based counter-memory, only ChatGPT, GPT-4, and PaLM2 over 60% probability of choosing parametric memory. For generation-based counter-memory, all models have more than 80% probability of choosing context knowledge.
Xu et al. (2023)	ChatGPT, GPT-4, Llama2, and Vicuna	Farm, BoolQ, TruthfulQA and NQ	In multiple rounds of dialogue, as the number of counter-memory context increases, the cumulative proportion of belief alteration of LLMs spans from 20.7% to 78.2%
		Inter-context conflict	
Jin et al. (2024a)	ChatGPT, Llama2, Baichuan2, FLAN-UL2 and FLAN-T5	NQ, TriviaQA, PopQA, and MuSiQue	When faced with conflicting evidence, ChatGPT's recall declined the least, but more than 10%.
Chen et al. (2023)	ChatGPT, ChatGLM, Vicuna, Qwen, and BELLE	RGB	As the noise in evidence increases, the performance of models will gradually decrease. When the noise rate exceeds 0.8, the performance of all models decreases by more than 20%.
Li et al. (2023a)	GPT-4, ChatGPT, PaLM2, and Llama2	CONTRADOC	Faced with self-contradictory documents, gpt4 has a more than 70% probability of determining the occurrence of a contradiction, while other models are less than 50%.
		Intra-memory conflict	
Mündler et al. (2023)	GPT-4, ChatGPT, Llama2, and Vicuna	MainTestSet	LLMs create contradictory content, with a probability of between 15.7% and 22.9%. More powerful models create fewer contradictory results.
Zhao et al. (2023b)	ChatGPT, GPT-4, Vicuna, and Llama2	FaVIQ, ComQA, GSM-8K, SVAMP, ARCChallenge, and CommonsenseQA	The findings of their research reveal that even GPT-4 can exhibit an inconsistency rate of 32% in FaVIQ.

Table 2: Comparison of quantitative results on the impact of various types of knowledge conflicts.

Reference	Model	Dataset	Quantitative Results
		Faithful to context	
Shi et al. (2023a)	Llama, OPT, GPT-Neo, and FLAN	NQ-SWAP, MemoTrap, and NQ	Their method improves GPT-Neo 20B by 54.4% on Memotrap and by 128% on NQ-SWAP where LLMs need to adhere to the given context.
Zhou et al. (2023d)	ChatGPT and Llama2	MRC and Re-TACRED	Compared to the zero-shot base prompts, their prompting method leads to a reduction of 32.2% for maintaining parametric knowledge for MRC and a 10.9% reduction for Re-TACRED on GPT-3.5. Similarly, on Llama2, there is a 39.4% reduction for MRC and a 57.3% reduction for Re-TACRED.
		Discriminating misinformation	
Hong et al. (2023)	ChatGPT and FiD	NQ and TQA	The authors train a discriminator with about 80% F1 score and use it to improve models performance above 5%.
Pan et al. (2023b)	ChatGPT	NQ-1500 and CovidNews	The author's mitigation method improves the accuracy by more than 10%.
		Disentangling sources	
Wang et al. (2023e)	ChatGPT	KNOWLEDGE CONFLICT	The authors' method achieved over 80% F1 score on contextual knowledge conflict detection.

Table 3: Comparison of quantitative results on the effectiveness of various mitigation strategies *w.r.t.* their objectives.