Resolving Knowledge Conflicts in Large Language Models

Yike Wang[*1]   **Shangbin Feng**[*2]   **Heng Wang**[3]   **Weijia Shi**[2]
Vidhisha Balachandran[4]   **Tianxing He**[2]   **Yulia Tsvetkov**[2]
[1]University of California, Berkeley   [2]University of Washington
[3]Xi'an Jiaotong University   [4]Carnegie Mellon University
yike_wang@berkeley.edu, shangbin@cs.washington.edu

Abstract

Large language models (LLMs) often encounter *knowledge conflicts*, scenarios where discrepancy arises between the internal parametric knowledge of LLMs and non-parametric information provided in the prompt context. In this work we ask *what are the desiderata for LLMs when a knowledge conflict arises and whether existing LLMs fulfill them*. We posit that LLMs should 1) *identify knowledge conflicts*, 2) *pinpoint conflicting information segments*, and 3) *provide distinct answers or viewpoints in conflicting scenarios*. To this end, we introduce an evaluation framework for simulating contextual knowledge conflicts and quantitatively evaluating to what extent LLMs achieve these goals. It includes diverse and complex situations of knowledge conflict, knowledge from diverse entities and domains, two synthetic conflict creation methods, and settings with progressively increasing difficulty to reflect realistic knowledge conflicts. Extensive experiments with the framework reveal that while LLMs perform well in identifying the existence of knowledge conflicts, they struggle to determine the specific conflicting knowledge and produce a response with distinct answers amidst conflicting information. To address these challenges, we propose new instruction-based approaches that augment LLMs to better achieve the three goals. Further analysis shows that abilities to tackle knowledge conflicts are greatly impacted by factors such as knowledge domain, while generating robust responses to knowledge conflict scenarios remains an open research question. Code and data are publicly available at github.com/yikee/Knowledge_Conflict.

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities to encode world knowledge (Peters et al., 2018; Petroni et al., 2019) and solve knowledge-intensive tasks (Roberts et al., 2020; Brown et al., 2020a). Nevertheless, their knowledge abilities are far from perfect (Sun et al., 2023; Hernandez et al., 2023; Muhlgay et al., 2023), leading to the emergence of knowledge augmentation approaches: using external sources (e.g., retrieval corpora (Fisch et al., 2019; Guu et al., 2020; Shi et al., 2023b), search engines (Press et al., 2022; Nakano et al., 2021), and other LMs (Feng et al., 2023d; Luo et al., 2023)) to provide relevant information in the prompt context. However, due to issues such as misinformation, varying perspectives, time-sensitive information, or knowledge updates, **knowledge conflicts** might arise, meaning that there is a discrepancy between *parametric knowledge* (the internal knowledge stored in LLM parameters) and *non-parametric knowledge* (the knowledge fetched from external sources (Chen et al., 2022; Xie et al., 2023)).

Prior research conducted preliminary studies by probing LLMs with knowledge conflicts and examined their behaviors in response (Chen et al., 2022; Li et al., 2023; Jin et al., 2024). The key findings are that LLMs' choices between knowledge sources, parametric or non-parametric, depend on factors including the coherence of the external knowledge (Xie et al., 2023) and model size (Longpre et al., 2021). This work extends these prior works by seeking

---

*equal contribution

a deeper understanding of whether LLMs can acknowledge knowledge conflicts and how they should respond. Specifically, we ask: *What should be the desirable behaviors of LLMs when knowledge conflicts arise?* and *Are LLMs currently exhibiting those desirable behaviors?* We argue that LLMs should not rely solely on either parametric or non-parametric information, but grant LLM users the agency to make informed decisions based on distinct answers (Floridi, 2023). In line with this goal, we hypothesize that LLMs should 1) identify the existence of knowledge conflicts, 2) pinpoint the specific information segments where knowledge conflicts occur, and 3) generate distinct responses based on all conflicting information. Achieving these desiderata, as shown in Figure 1, enables LLMs to not only acknowledge the existence of knowledge conflicts but also navigate them skillfully, resulting in responses that are more accurate, comprehensive, and, ultimately, trustworthy.

To this end, we introduce a framework to simulate contextual knowledge conflicts and evaluate whether LLM's behavior aligns with the three desiderata. Specifically, we curate a list of 10k entities covering 200 subject areas, while employing two techniques to generate synthetic knowledge conflicts tailored to a specific context. We establish three distinct tasks with increasing complexity to reflect the three goals: 1) *Contextual Knowledge Conflict Detection*: identify the presence of knowledge conflicts, 2) *QA-Span Knowledge Conflict Detection*: determine whether there is a knowledge conflict specifically in a span which is relevant to the question; and 3) *Distinct Answers Generation*: provide distinct answers by leveraging all pieces of conflicting information. They focus on different aspects of conflict-handling abilities and together serve as a comprehensive evaluation protocol.
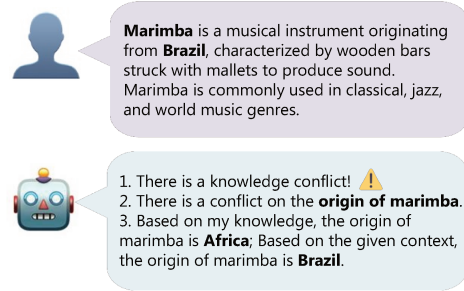

Figure 1: We expect LLMs to 1) acknowledge knowledge conflicts, 2) point out the specific conflicting segments, and 3) generate different answers based on conflicting pieces of information.

With extensive experiments, we find that while LLMs perform well above random on Task 1, identifying the existence of knowledge conflicts within contextual information, they encounter notable challenges when it comes to Tasks 2 and 3, which require LLMs to precisely pinpoint these conflicts and provide distinct answers given conflicting context. To address these challenges, we further propose new instruction-based approaches to reflect a wide array of reasoning properties, such as decomposing tasks, breaking down context passages, localization, and more. Through these approaches, we successfully enhance the performance of GPT-3.5-TURBO on Task 1 and Task 3, improving LLM's abilities to acknowledge knowledge conflicts and generate distinct answers amidst conflicting information. Further analyses demonstrate that factors such as knowledge domain are relevant, while robust handling of knowledge conflicts remains an open research question.

2 Framework

We present a framework (Figure 2), leveraging diverse knowledge sources, various conflict creation methods, and progressively challenging settings to reflect real-world knowledge conflicts and quantitatively assess LLMs' capability to address knowledge conflicts.

2.1 Knowledge Scope

We generate an entity list as the starting point by prompting LLMs in a zero-shot manner: we instruct the LLM to return 20 distinct domains such as Computer Science, accompanied by 10 fields within each domain such as Artificial Intelligence and Human-Computer Interaction, and then 50 entities specific to each field such as Neural networks and User Interface. As a result, we obtain 9,083 unique entities after filtering out duplicates, covering diverse knowledge areas across various domains. We utilize the generated entity list instead of other publicly accessible entity lists (Pellissier Tanon et al., 2020; Heist & Paulheim, 2020), so it is highly likely that LLMs are familiar with these entities and would contain knowledge and information about them. Note that the framework is independent of the entity list, thus our approach could be easily extended to other domains, subject areas, and more.
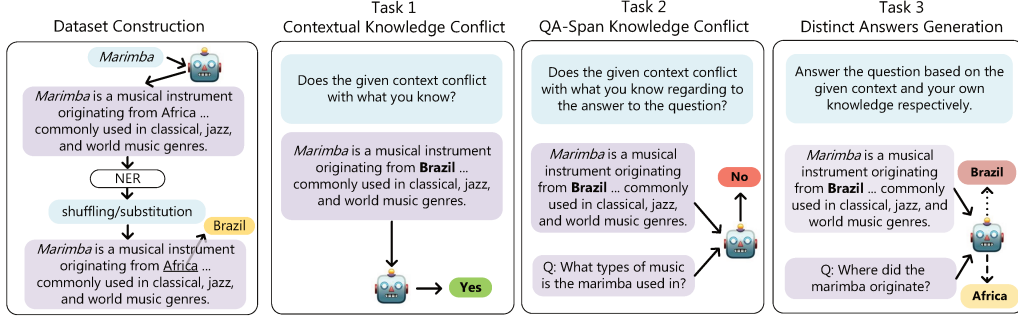
Figure 2: We introduce a framework to comprehensively analyze and improve LLMs' handling of knowledge conflicts. The framework covers concrete spans where knowledge conflicts arise, and facilitates meaningful outputs, granting users the agency to find appropriate responses among conflicting information.

2.2 Knowledge Conflict Generation

For each entity, we create two pieces of information by first eliciting the LLM's parametric knowledge about the entity, and then factually modifying it to construct a conflicting knowledge to later put into the prompt context, such that there is a knowledge conflict between these two contexts. We detail the methodology below.

Parametric Knowledge Elicitation   We instruct LLMs to produce contextual information about each entity under a closed-book setting with the prompt *"Give me some context about {entity} in 50 words."* In this case, LLMs rely solely on their internal parametric knowledge, devoid of external evidence, to generate the requested context. As a result, we adopt the generated context as its parametric knowledge. Different parametric knowledge needs to be elicited when experimenting with different LLMs.

Conflicting Knowledge Creation

- *In-domain Named Entity Substitution*: Inspired by previous works that effectively utilize the entity substitution method (Longpre et al., 2021; Xie et al., 2023; Liu et al., 2024), We employ spaCy's NER models (Honnibal et al., 2020; Liu et al., 2019) to identify named entity categorized as *ordinal, cardinal, date, person, organization,* and *location*. We randomly select an identified entity and perform substitution: all occurrences of the selected entity are substituted with another entity of the same type drawn from an in-domain corpus, i.e., an entity of the type "person" will be substituted with another entity of the type "person" found in the parametric knowledge contexts from the same domain.

- *In-domain Entity Shuffling*: For cases where the NER models fail to identify entities of the six categories, we proceed to shuffle their "main" entities – entities in Section 2.1 that we used to generate these contexts. Concretely, we replace all occurrences of the main entity in the context with another main entity in a context from the same domain.

2.3 Tasks

After obtaining pairs of passages that are in conflict with each other, we create three tasks to examine LLM's ability to 1) recognize the existence of knowledge conflicts, 2) pinpoint conflicting information segments, and 3) provide different answers to each of the conflicting passages.

Task 1: Contextual Knowledge Conflict   We set up a binary classification task in which a single piece of context, either its parametric knowledge or the conflicting knowledge, and the instruction *"Does the given context conflict with what you know? Yes/No"* are given in the prompt. The answer *"Yes"* is expected when the conflicting knowledge is given and *"No"* in case of the parametric knowledge. We use Precision, Recall, and F1 as evaluation metrics.

Task 2: QA-Span Knowledge Conflict   It is often the case that not all pieces of information within a passage are conflicting, so in addition to detecting overall contextual knowledge

conflict (Task 1), it is crucial for LLMs to pinpoint the specific piece of information where these conflicts arise. We instruct TEXT-DAVINCI-003 (OpenAI, 2022) with the prompt *"Given the context, generate a question to which the only single answer is the word {entity} (the question should not contain the word {entity})"*, where the *"entity"* is the entity substituted or shuffled in the conflict generation step, and the conflicting context in a zero-shot manner to generate a question asking about the conflicting segments of the conflicting context. The prompt *"Given the context, generate a question unrelated to {entity}"* is employed to generate a question asking about the non-conflicting segments of the conflicting context. For each knowledge pair, we generate one question related to the conflicting knowledge segments and one question to the non-conflicting knowledge segments. We prompt the LLM with a single context (either parametric knowledge or conflicting knowledge) and a single question (either question about the conflicting segments or question about the non-conflicting segments) with the instruction *"Does the given context conflict with what you know regarding the answer to the question? Yes/No"* for a binary classification. The positive answer is only expected when the conflicting knowledge and the question about the conflicting segments are given. Though we can assess LLMs directly by letting them to identify conflicting segments, we opt for this QA-based method which aligns better with real-world scenarios where users ask questions that might not rely on the conflicting segments. Again, we employ Precision, Recall, and F1 for evaluation.

Task 3: Distinct Answers Generation   While previous studies (Longpre et al., 2021; Xie et al., 2023; Mallen et al., 2023) explored various factors that impact the LLMs to choose between their parametric knowledge and external sources, we believe that it is important to defer agency to the users, i.e., in the face of ambiguity and knowledge conflicts, LLMs should return multiple answers and let users make the choices. Therefore, we test LLMs' ability to generate different answers given conflicting contexts in this task. Specifically, the LLM is given the conflicting text and the question about the conflicting segments of text along with the prompt *"Answer the question based on the given context and your own knowledge respectively."* The ground truths would be the answer based on the conflicting passage and the answer generated by LLMs when only the question is presented in a zero-shot manner. We evaluate the accuracy of parametric-based answers, the accuracy of conflicting-knowledge-based answers, and the accuracy of simultaneously providing both answers.

2.4   Dataset Analysis

It's worth noting that our conflict generation methods may fail under situations where the context is not unique to a specific entity, for example, when there are multiple individuals holding the title of "chief scientist." To further validate the effectiveness of our conflict generation techniques, we conduct human evaluations for Task 1 and Task 2. Results show that 96% of Task 1 and 83.5% of Task 2 contain perfectly clear knowledge conflicts. The Fleiss' Kappa (Fleiss, 1971) among the five annotators is 0.51, indicating moderate agreement. A detailed breakdown of the dataset can be found in Table 1.

Domain	Substitution	Shuffling	Total
Mathematics	220	273	493
Biology	107	337	444
Chemistry	53	434	487
Physics	145	323	468
Psychology	59	377	436
Computer Science	60	420	480
History	449	2	451
Literature	340	137	477
Sociology	41	385	426
Economics	67	377	444
Political Science	253	202	455
Philosophy	108	293	401
Anthropology	106	338	444
Geology	200	260	460
Linguistics	47	380	427
Art History	426	59	485
Environmental Science	68	363	431
Health Sciences	101	377	478
Communications	55	392	447
Music	175	274	449
Total	3,080	6,003	9,083

Table 1: The dataset we construct through our framework comprises 9,083 distinct entities that are approximately evenly distributed across 20 different domains. Around one-third of the instances of knowledge conflicts stem from named entity substitution, while the remaining two-thirds result from entity shuffling.

3   Experiment Settings

3.1   Baselines

We evaluate prominent LLM prompting approaches including zero-shot (Liu et al., 2021b), few-shot (Brown et al., 2020b), Chain-of-Thought (CoT) (Wei

et al., 2022), Generated Knowledge Prompting (GKP) (Liu et al., 2021a), Self-ask (Press et al., 2022), Break-down, and Self-Consistency (SC) (Wang et al., 2023b) with our framework. See details in Appendix B.

3.2 Models and Settings

We use ChatGPT (OpenAI, 2022) (GPT-3.5-TURBO) as the main LLM in the experiments unless otherwise stated. For Self-Consistency that requires multiple samples for a problem, we sample 5 responses with temperature $\tau = 0.7$ following (Wang et al., 2023b); for all the other experiments, we use temperature $\tau = 0$. For few-shot prompting approaches, the input prompt includes four in-context exemplars and their solutions before the problem of interest under Task 1 and Task 2, and two such pairs under Task 3. The in-context exemplars are drawn from different primary fields and different conflict generation methods, and balanced between positive and negative samples.

4 Results

Task 1: Contextual Knowledge Conflict Table 2 shows that on Task 1, LLMs display a tendency to declare "*no conflict*", which results in a nearly perfect precision but a low recall. This inclination toward asserting the absence of conflicts can raise doubts about negative predictions, as it doesn't necessarily indicate a genuine assessment of the absence of conflicts. However, these concerns are alleviated when considering the results obtained using CoT, where providing reasons is obligatory. In this scenario, both GKP and Self-ask, methods that rely on explicit classification, do not yield strong performance. This indicates that accurately identifying contextual knowledge conflicts through explicit means is not a trivial task. Overall, the LLM demonstrates an above-random ability in contextual knowledge conflict identification.

Method	Prec.	Rec.	F1
ZERO-SHOT	**0.999**	0.144	0.251
FEW-SHOT	**0.999**	0.351	0.520
CoT	<u>0.998</u>	0.639	0.779
CoT + SC	<u>0.998</u>	0.644	<u>0.783</u>
GKP + SC	**0.999**	0.475	0.644
SELF-ASK	0.995	0.486	0.653
BREAK-DOWN	0.863	<u>0.693</u>	0.768
Ours	0.893	**0.728**	**0.802**

Table 2: Performance on Task 1: Contextual Knowledge Conflict Detection. The best results are **bold-faced** and the second-best ones are <u>underlined</u>. Our proposed approach outperforms all baselines on F1.

Task 2: QA-Span Knowledge Conflict Table 3 shows that when the LLM is tasked with the precise identification of knowledge conflicts, its performance experiences a considerable decline. Among all the baseline methods, the self-ask prompting approach stands out as the most effective, and we find that the generated intermediate answers help to narrow down the scope of knowledge conflicts and encourage localization. Also, we observe a consistent pattern where precision exceeds recall, which revalidates LLMs' tendency to assert "*no conflict*". Overall, LLMs struggle to precisely pinpoint the piece of information where these conflicts arise.

Method	Prec.	Rec.	F1
ZERO-SHOT	0.615	0.151	0.242
FEW-SHOT	0.395	**0.860**	<u>0.541</u>
CoT	0.843	0.375	0.519
CoT + SC	<u>0.875</u>	0.367	0.517
GKP + SC	0.508	<u>0.499</u>	0.504
SELF-ASK	**0.898**	0.474	**0.621**
BREAK-DOWN	0.614	0.413	0.494
Ours	0.718	0.426	0.535

Table 3: Performance on Task 2: QA-Span Knowledge Conflict Detection. The best results are **bold-faced** and the second-best ones are <u>underlined</u>. Self-ask prompting stands out as the strongest baseline method.

Task 3: Distinct Answers Generation Table 4 shows the results under Task 3 where LLMs are directed to provide answers based on the non-parametric context and its parametric knowledge concurrently. Across all the prompting methods, except for zero-shot where only a single answer is returned in most cases, the accuracy of answers based on the conflicting knowledge surpasses that of parametric-based answers. Break-down prompting is not applicable in this task, and we have omitted Self-Consistency due to the limited improvements it offers in the first two tasks and cost considerations. Overall, the LLM struggles to provide distinct answers simultaneously with the accuracy of getting both correct hovering around 50%, requiring further research and exploration.

5 Proposed Approach

Recent studies ([Shi et al., 2023a](); [Wang et al., 2023a]()) have shown that instruction-based methods work well to induce new abilities in LLMs. We additionally explore and propose new instruction-based approaches to investigate whether instructions tailored to the context of knowledge conflicts would improve upon generic approaches. Full prompt text is presented in Appendix E.

Method	Para.	Conflicting.	Both
ZERO-SHOT	0.400	0.350	0.031
FEW-SHOT	0.372	0.765	0.285
COT	0.575	0.782	0.473
GKP	0.643	0.814	0.551
SELF-ASK	0.611	0.735	0.464
Ours	**0.658**	**0.815**	**0.569**

Table 4: Performance on Task 3: Distinct Answers Generation. The best results are **bold-faced** and the second-best ones are underlined. Our approach enables LLMs to generate distinct answers supported by different knowledge sources respectively.

Task 1: Contextual Knowledge Conflict We propose to employ a four-step approach: 1) elicit knowledge about the main entity, 2) break down the entire context into individual sentences, which draws LLMs' attention to every single detail, and identify sentences that can be verified by the knowledge elicited in step 1, 3) classify whether these sentences conflict with the knowledge elicited in a sentence-by-sentence manner, and 4) classify whether the remaining sentences conflict with its parametric knowledge (using all the internal knowledge in addition to the knowledge elicited in step 1). For steps 2), 3), and 4), a localization procedure is included, which means apart from returning the answer, the LLMs also need to return their reasoning steps. The main idea is that we promote fine-grained analysis into the sentence-level so that the LLM could better classify and attend to those parts, leaving all the *vague* sentences to the final step. Table 2 shows that our proposed method exhibits a higher F1 score compared to all the baseline methods, albeit with a slight reduction in precision. The efficacy of both the Break-down baseline and our proposed approach underscores that the capacity to discern contextual knowledge conflicts is contingent upon the context's length.

Task 2: QA-Span Knowledge Conflict Similarly, we propose to break down the task and fine-grain the context into sentences that can be used to answer the given question. Specifically, the LLMs are instructed to 1) disregard the given context, answer the given question solely based on their own beliefs, 2) among the given context, find sentences that can be used to answer the given question: if such sentences exist, extract answers from the sentences and determine whether these answers conflict with the answers generated in step 1; if no such sentences exist, report that there is no conflict. As shown in Table 3, unfortunately, our approach falls short of outperforming all the baseline methods in this setting after great exploration, indicating that instruction-based approaches might be limited in this scenario. This opens up opportunities for future research to enhance the capability of LLMs in pinpointing instances of knowledge conflicts.

Task 3: Distinct Answers Generation In order to get more accurate parametric-knowledge-based answers and conflicting-knowledge-based answers, we propose to include "keywords" such as "*solely*" and "*disregard*" to separate the two knowledge sources apart. Also, after generating the response based on one knowledge source, we instruct the LLMs to repeat the question again as LLMs have exhibited limited capability in retaining information across extended contextual spans ([Chen et al., 2023](); [Liu et al., 2023]()). Table 4 shows that our proposed method attains superior performance across all three metrics.

6 Analysis

Breakdown by Factors To examine the factors that may influence the ability of LLMs to identify contextual knowledge conflicts, pinpoint knowledge conflicts, and offer distinct answers when confronted with conflicting information sources, we delve deeper into our results by categorizing them into various domains and conflict generation methods. We also put forward hypotheses regarding the potential reasons for the effects these factors have.
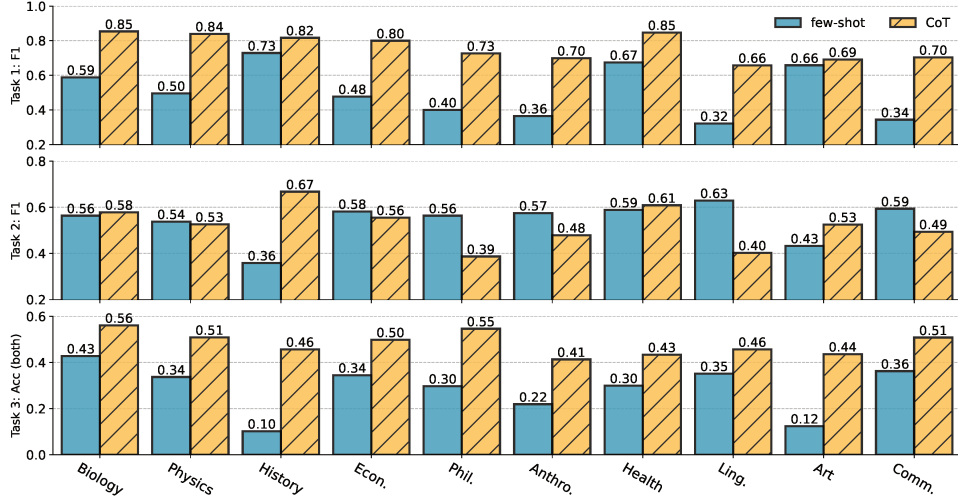
Figure 3: Model performance across knowledge domains on three different tasks. The factor of knowledge domain has a substantial effect on LLM performance on all three tasks, while certain domains (e.g. `art history` and `anthropology`) pose greater challenges.

Prompt	Task 1		Task 2	
	Few-shot	CoT	Few-shot	CoT
Does the given context conflict with what you know...?	0.529	0.805	0.541	0.546
Does the provided information contradict your existing knowledge...?	0.529	0.795	0.541	0.580
Is the given information consistent with what you know...?	0.734	0.784	0.143	0.671
Based on your knowledge, is there anything wrong with the given information...?	0.616	0.810	0.442	0.628
standard deviation	0.084	0.010	0.163	0.047

Table 5: F1 scores on Tasks 1 and 2 when using different instructions. CoT is more robust than FEW-SHOT in the face of minor changes in instruction texts.

- *Knowledge Domain*: As shown in Figure 3, LLMs exhibit a higher proficiency in recognizing (Task 1) and pinpointing (Task 2) contextual knowledge conflict within the domains of `History` and `Health Sciences`. Regarding Task 3, providing distinct answers, LLMs excel in the domains of `Biology`. These results show varying proficiency in handling knowledge conflicts across knowledge domains: we hypothesize that it has to do with the quantity of conflicting information present in the pre-training data. Essentially, if LLMs encounter a substantial amount of conflicting information during their pre-training within a specific domain, they tend to perform better in that particular domain. We leave the results for the other 10 knowledge domains in Appendix D.

- *Conflict Generation Method*: Figure 4 demonstrates that when we dissect the results according to the two synthetic methods used to generate conflicts (i.e., In-domain Named Entity Substitution and In-domain Entity Shuffling), it becomes evident that LLMs exhibit enhanced performance in scenarios where conflicts are induced through entity shuffling. This outcome aligns with intuition since LLMs find it more straightforward to identify and specify knowledge conflicts when the conflict pertains to the main entity. Results for Task 3 can be found in Appendix D.

Prompt Consistency Check   LLMs might be sensitive to subtle changes in the prompt text (Zhu et al., 2023). To further examine the potential impact of instruction phrasing on LLMs' performance in given tasks, we introduce three additional formulations of instructions for each task. Table 5 reveals that, on the whole, performance remains stable across various phrasings. The results also indicate that the CoT prompting method enhances prompt robustness, as evidenced by significantly lower standard deviations in comparison to the few-shot prompting method. See Appendix D for prompt consistency check on Task 3.
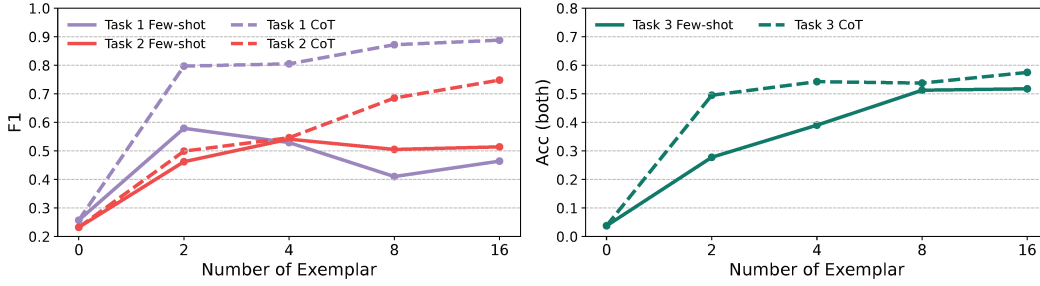
Figure 5: Model performance with various numbers of in-context exemplars. CoT benefits more from increasing exemplars, while their impact quickly saturates at 8 examples.

Method	Para.		Conflicting.		Both	
	GPT-4	change w/ turbo	GPT-4	change w/ turbo	GPT-4	change w/ turbo
ZERO-SHOT	0.206	-0.194	0.628	+0.278	0.057	+0.026
FEW-SHOT	0.638	+0.266	0.923	+0.158	0.604	+0.319
CoT	0.691	+0.116	0.877	+0.094	0.625	+0.153
GKP	0.723	+0.080	0.865	+0.051	0.656	+0.105
SELF-ASK	0.684	+0.073	0.880	+0.145	0.619	+0.155

Table 6: Performance with GPT-4 as the base model and changes compared to GPT-3.5-TURBO. GPT-4 exhibits improvements across all baselines, while it still falls short of optimal.

Number of In-context Exemplars   We explore whether the quantity of in-context exemplars affects the ability to tackle knowledge conflicts. Specifically, we change the number of in-context exemplars on a subset of the dataset across the three tasks. As shown in Figure 5, we generally observe that the F1 score plateaus when there are as few as two in-context exemplars. However, including additional exemplars proves beneficial in the case of CoT prompting for Task 2 and few-shot prompting for Task 3. To sum up, increasing the number of exemplars may improve LLMs' capability of handling knowledge conflicts, but its effect is limited and far from inducing perfectly robust approaches.
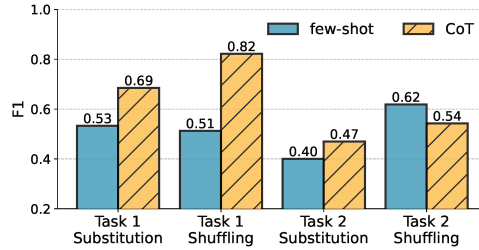

Figure 4: Performance on Tasks 1 and 2 when two conflict creation strategies, named entity substitution and entity shuffling, are separately employed. LLMs demonstrate better performance when conflicts are created by in-domain entity shuffling.

More Capable LLMs   We also investigate the competence of other LLMs in tackling knowledge conflicts, which encompass GPT-4 (OpenAI, 2023), in Table 6. Due to budget constraints, we only assess its performance on the most challenging Task 3. As an LLM trained on an unprecedented scale of compute and data, GPT-4 showcases increased accuracy in generating both parametric-based answers and conflicting-knowledge-based answers. However, its performance has yet to reach an optimal level, indicating that mere scaling does not solve the challenge of knowledge conflicts.

Finetuning   We also finetune the ChatGPT (GPT-3.5-TURBO) under Tasks 1 and 2, with training data from the domain of `Political Science` and testing data from the domain of `Geology` to avoid the chance of overfitting. As shown in Table 7, fine-tuning is exceptionally beneficial for Task 1, leading to an impressive F1 score of 0.953. However, in the case of Task 2, the impact of fine-tuning is less pronounced, resulting in a comparatively modest F1 score of 0.745. An intriguing observation from Task 2 is that the recall exceeds the precision, which implies a noteworthy reduction in the model's tendency to assert negativity in acknowledging knowledge conflicts.

7 Related Work

Understanding and expanding the knowledge abilities of LLMs Previous works have demonstrated that LLMs have incorporated factual knowledge within their parameters and exhibit considerable potential in recalling factual information (Peters et al., 2018; Petroni et al., 2019; Yu et al., 2023; Mruthyunjaya et al., 2023). However, existing research also reveals that their inherent knowledge is not without flaws (Wu et al., 2022; Pan et al., 2023): outdated knowledge

Task	Method	Prec.	Rec.	F1
TASK 1	fine-tune	0.977	0.930	0.953
	prev.best	0.999	0.728	0.802
TASK 2	fine-tune	0.689	0.811	0.745
	prev.best	0.898	0.860	0.621

Table 7: Fine-tuning results on Tasks 1 and 2. While the performance surpasses baselines, it remains imperfect.

(Hernandez et al., 2023; Yu & Ji, 2023; Padmanabhan et al., 2023), factuality issues (Lee et al., 2022; Feng et al., 2023b), hallucination (Ji et al., 2023; Zhang et al., 2023), and more are common challenges in LLM knowledge abilities. In response, researchers enhance these capabilities through approaches such as retrieval augmentation (Lewis et al., 2020; Guu et al., 2020; Borgeaud et al., 2022; Shi et al., 2023b; Jiang et al., 2023; Zhang et al.), search engine integration (Nakano et al., 2021; Press et al., 2022; Feng et al., 2023a), and incorporating other neural LMs (Feng et al., 2023d; Luo et al., 2023; Du et al., 2023). In this work, we specifically focus on the issue of *knowledge conflict*, when there is a conflict between internal parametric knowledge and external non-parametric knowledge. Without a thorough understanding of how LLMs react to and manage knowledge conflicts, the reliability of their responses may come into question.

Knowledge Conflict in LLMs Previous work on knowledge conflicts primarily focuses on factors impacting models' choice between parametric knowledge and non-parametric knowledge under QA settings. Mallen et al. (2023) finds that conflicting memories are effective for less popular facts; Longpre et al. (2021) explores the effects of model size and retrieval quality by identifying QA instances with named entity answers and substituting mentions of the entity in the gold document with an alternate entity, thus changing the answer; Xie et al. (2023) reveals that when both supportive and contradictory evidence to their parametric memory are present, LLMs show a strong confirmation bias and tend to cling to their parametric memory; Neeman et al. (2023) investigates training data augmentation methods to disentangle parametric and contextual knowledge with counterfactual question answering; Qian et al. (2023) employs the idea of knowledge graphs and systematically generates distractors with various methods, degrees, positions, and formats, showing that LLMs tend to produce responses that deviate from their parametric knowledge, particularly with direct conflicts. Nevertheless, an intriguing and underexplored aspect is to rethink the desiderata for LLMs when confronted with knowledge conflicts, and whether their current responses align with such objectives. We argue that LLMs should 1) *identify knowledge conflicts*, 2) *pinpoint conflicting information segments*, and 3) *provide distinct answers in conflicting scenarios*. We propose a framework and conduct extensive experiments to evaluate and improve LLMs' ability to tackle knowledge conflicts.

8 Conclusion

We introduce an evaluation framework to simulate contextual knowledge conflicts and quantitatively evaluate LLMs' ability to 1) identify contextual knowledge conflicts, 2) pinpoint conflicting knowledge segments, and 3) provide distinct answers or viewpoints amidst conflicts. Extensive experiments demonstrate that LLMs excel at simply identifying knowledge conflicts, but struggle with fine-grained analysis and providing distinct responses. We propose instruction-based approaches that successfully improve in Task 1 and Task 3, while challenges persist in all tasks, especially in Task 2 of pinpointing conflicts. Further analyses show that factors like exemplars, domains, and conflict simulation methods greatly impact LLM's ability to tackle knowledge conflicts. Our comprehensive framework and in-depth study offer a comprehensive understanding of whether existing LLMs could generate desirable responses amid knowledge conflicts and provide quantitative avenues for evaluating and improving the ability to tackle knowledge conflicts.

Limitations

Conflicting knowledge may not be factually correct. In line with the prior work (Balachandran et al., 2022), we replace facts with a randomly related entity for simplicity. While future work could focus on generating more plausible knowledge alternatives to better align with real-world scenarios, our primary emphasis is the existence of knowledge conflicts instead of factual correctness.

Assume the top answer as the unique parametric knowledge. Throughout this paper, we assume the top answer generated with prompting in a zero-shot manner as the unique *parametric knowledge* of the LLM to the given entity or question (exploration on the possibility of multiple parametric knowledge answers are in Appendix F). However, there is a chance of hallucination in those generated contexts (Ji et al., 2023; Zhang et al., 2023). We argue that since our entity list is generated by probing LLMs, there is a bias towards entities well-known to LLMs, reducing the likelihood of hallucination. The population bias, which is believed to impact whether LLMs rely on parametric knowledge or counter-parametric knowledge (Mallen et al., 2023), is not a significant factor as the LLMs are not instructed to make such a choice in our study. Moreover, there is an ongoing debate about whether hallucination should be considered a form of parametric knowledge, and our focus is not on whether the parametric knowledge is factually accurate or not.

Real-world knowledge conflicts might be more complex. While we employ multiple conflict generation methods and settings to simulate knowledge conflicts, it's crucial to note that our dataset is synthetically generated and limited to word-level edits. Knowledge conflicts found in real-world data might introduce a higher level of complexity (see experiments on multi-hop questions and subjective datasets in Appendix G and H). For instance, they may involve entirely new entities, increasing the risk of hallucination, while posing new challenges to LLMs.

Acknowledgments

Ethics Statement

The problem of knowledge conflicts has great ethical implications. In situations where there exists a knowledge conflict, a disparity between external non-parametric knowledge and internal parametric knowledge, there is typically no inherent inclination to favor one over the other. Decisions regarding which knowledge source to trust are context-dependent. For example, in cases necessitating knowledge updates, external knowledge may take precedence. Conversely, in instances involving potential malicious interference, reliance on internal parametric beliefs is advisable. However, when the external information pertains to opinions or diverse perspectives, there may not be an unequivocal correct answer at all. Consequently, navigating knowledge conflicts in such instances raises notable ethical considerations. It is essential to recognize that the baselines and proposed approaches in this study, while valuable for research purposes, are far from perfect and may not perfectly mirror real-world complexities and ethical dilemmas associated with knowledge conflicts.

References

Vidhisha Balachandran, Hannaneh Hajishirzi, William Cohen, and Yulia Tsvetkov. Correcting diverse factual errors in abstractive summarization via post-editing and language model infilling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9818–9830, 2022.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020a.

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

Hung-Ting Chen, Michael Zhang, and Eunsol Choi. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2292–2307, 2022.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023.

Jiazhan Feng, Chongyang Tao, Xiubo Geng, Tao Shen, Can Xu, Guodong Long, Dongyan Zhao, and Daxin Jiang. Knowledge refinement via interaction between search engines and large language models. *arXiv preprint arXiv:2305.07402*, 2023a.

Shangbin Feng, Vidhisha Balachandran, Yuyang Bai, and Yulia Tsvetkov. Factkb: Generalizable factuality evaluation using language models enhanced with factual knowledge. *arXiv preprint arXiv:2305.08281*, 2023b.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11737–11762, July 2023c.

Shangbin Feng, Weijia Shi, Yuyang Bai, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. Cook: Empowering general-purpose language models with modular and collaborative knowledge. *arXiv preprint arXiv:2305.09955*, 2023d.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. Mrqa 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pp. 1–13, 2019.

Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382, 1971.

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

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

Nicolas Heist and Heiko Paulheim. Entity extraction from wikipedia list pages. In Andreas Harth, Sabrina Kirrane, Axel-Cyrille Ngonga Ngomo, Heiko Paulheim, Anisa Rula, Anna Lisa Gentile, Peter Haase, and Michael Cochez (eds.), *The Semantic Web*, pp. 327–342, Cham, 2020. Springer International Publishing. ISBN 978-3-030-49461-2.

Evan Hernandez, Belinda Z Li, and Jacob Andreas. Measuring and manipulating knowledge representations in language models. *arXiv preprint arXiv:2304.00740*, 2023.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, et al. spacy: Industrial-strength natural language processing in python. 2020.

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

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. *arXiv preprint arXiv:2305.06983*, 2023.

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

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

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

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. Large language models with controllable working memory. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 1774–1793, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.112. URL https://aclanthology.org/2023.findings-acl.112.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. Generated knowledge prompting for commonsense reasoning. In *Annual Meeting of the Association for Computational Linguistics*, 2021a.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*, 2023.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55:1 – 35, 2021b.

Yantao Liu, Zijun Yao, Xin Lv, Yuchen Fan, Shulin Cao, Jifan Yu, Lei Hou, and Juanzi Li. Untangle the knot: Interweaving conflicting knowledge and reasoning skills in large language models. *arXiv preprint arXiv:2404.03577*, 2024.

Yinhan Liu, Myle Ott, and Naman Goyal. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

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

Ziyang Luo, Can Xu, Pu Zhao, Xiubo Geng, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. Augmented large language models with parametric knowledge guiding. *arXiv preprint arXiv:2305.04757*, 2023.

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

Vishwas Mruthyunjaya, Pouya Pezeshkpour, Estevam Hruschka, and Nikita Bhutani. Re-thinking language models as symbolic knowledge graphs. *ArXiv*, abs/2308.13676, 2023.

Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov, Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. Generating benchmarks for factuality evaluation of language models. *arXiv preprint arXiv:2307.06908*, 2023.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. DisentQA: Disentangling parametric and contextual knowledge with counterfactual question answering. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10056–10070, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.559. URL https://aclanthology.org/2023.acl-long.559.

OpenAI. Chatgpt. 2022. URL https://openai.com/blog/chatgpt.

OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023.

Shankar Padmanabhan, Yasumasa Onoe, Michael JQ Zhang, Greg Durrett, and Eunsol Choi. Propagating knowledge updates to lms through distillation. *arXiv preprint arXiv:2306.09306*, 2023.

Xiaoman Pan, Wenlin Yao, Hongming Zhang, Dian Yu, Dong Yu, and Jianshu Chen. Knowledge-in-context: Towards knowledgeable semi-parametric language models. In *The Eleventh International Conference on Learning Representations*, 2023.

Thomas Pellissier Tanon, Gerhard Weikum, and Fabian Suchanek. Yago 4: A reason-able knowledge base. In *The Semantic Web: 17th International Conference, ESWC 2020, Heraklion, Crete, Greece, May 31–June 4, 2020, Proceedings 17*, pp. 583–596. Springer, 2020.

Matthew E Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1499–1509, 2018.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 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)*, pp. 2463–2473, 2019.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. *ArXiv*, abs/2210.03350, 2022.

Cheng Qian, Xinran Zhao, and Sherry Tongshuang Wu. "merge conflicts!" exploring the impacts of external distractors to parametric knowledge graphs. *ArXiv*, abs/2309.08594, 2023. URL https://api.semanticscholar.org/CorpusID:261875641.

Adam Roberts, Colin Raffel, and Noam Shazeer. 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)*, pp. 5418–5426, Online, November 2020.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pp. 31210–31227. PMLR, 2023a.

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

Taylor Sorensen, Liwei Jiang, Jena D Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, et al. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19937–19947, 2024.

Kai Sun, Y. Xu, Hanwen Zha, Yue Liu, and Xinhsuai Dong. Head-to-tail: How knowledgeable are large language models (llm)? a.k.a. will llms replace knowledge graphs? *ArXiv*, abs/2308.10168, 2023.

Heng Wang, Shangbin Feng, Tianxing He, Zhaoxuan Tan, Xiaochuang Han, and Yulia Tsvetkov. Can language models solve graph problems in natural language? *arXiv preprint arXiv:2305.10037*, 2023a.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022.

Yuhuai Wu, Markus Norman Rabe, DeLesley Hutchins, and Christian Szegedy. Memorizing transformers. In *International Conference on Learning Representations*, 2022.

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

Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-Li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, et al. Kola: Carefully benchmarking world knowledge of large language models. *arXiv preprint arXiv:2306.09296*, 2023.

Pengfei Yu and Heng Ji. Self information update for large language models through mitigating exposure bias. *arXiv preprint arXiv:2305.18582*, 2023.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023.

Yunxiang Zhang, Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. Merging generated and retrieved knowledge for open-domain qa.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Weirong Ye, Neil Zhenqiang Gong, Yue Zhang, and Xingxu Xie. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *ArXiv*, abs/2306.04528, 2023.

A   Discussion and Future Work

In our work, we argue that LLMs should be able to acknowledge knowledge conflicts, pinpoint conflicting information segments, and subsequently provide distinct answers grounded on different pieces of conflicting information. This capability not only enhances the robustness of LLMs against malicious attacks by reducing their reliance on a single information source, but also addresses ethical concerns by giving agency to LLM users. Tackling knowledge conflicts in LLMs also holds significance in a socio-political context, as achieving these goals can help LLMs better reflect the plurality of opinions when it comes to social and political issues (Feng et al., 2023c).

In our current task setup, there is one answer derived from the provided context, along with another answer drawn from the parametric knowledge. However, future research can delve into evaluating how well LLMs perform in scenarios where questions can have multiple valid answers stemming from either external information or internal beliefs. Additionally, it would be interesting to explore the possibility of amalgamating the three distinct tasks into one coherent challenge (i.e., instruct LLMs to identify knowledge conflicts, pinpoint knowledge conflicts, and generate distinct answers at the same time) and see if that helps with the performance on individual tasks. Furthermore, it could be intriguing to assess how well LLMs handle perspectives and opinions as conflicting information in addition to facts and knowledge and how the findings might shift with different LLMs, especially those that are open-source.

B   Prompting Methods

Zero-shot prompting   (Liu et al., 2021b) presents LLMs with a problem statement and asks for a direct answer, without any exemplars or intermediate reasoning steps.

Few-shot prompting   (Brown et al., 2020b) leverages a few exemplars, pairs of problems and answers, to prompt in-context learning in LLMs.

Chain-of-Thought prompting (CoT)   (Wei et al., 2022) includes a reasoning path in in-context exemplars and guides LLMs to follow similar reasoning steps to reach an answer. In Task 1, we guide LLMs to deconstruct the given context into atomic facts and check if the number of inconsistencies is greater than zero. In Tasks 2 and 3, we lead LLMs to generate the answers based on parametric knowledge and the answers based on the given context separately before the final response.

Generated Knowledge Prompting (GKP)   (Liu et al., 2021a) involves extracting knowledge from LLMs, and providing it as an additional input when answering a question. We elicit LLMs' parametric knowledge about the main entity in the given context as the supplementary input.

Self-ask prompting   (Press et al., 2022) requires LLMs to explicitly formulate the next follow-up question they should inquire before answering it. We employ this approach to generate questions on parametric knowledge and provided context.

Break-down prompting   guides LLMs to solve problems or answer questions at the sentence level, and then integrates all responses in the end. We instruct LLMs to perform classification on a sentence-by-sentence basis and then consolidate these individual responses into a coherent answer.

Self-Consistency (SC)   (Wang et al., 2023b) is a decoding strategy that samples a diverse set of reasoning paths and selects the most consistent answer by marginalizing out the sampled reasoning paths, leveraging the idea that a complex reasoning problem typically admits multiple different ways of thinking leading to its unique correct answer. In our experiments, Self-Consistency is used in conjunction with CoT and GKP.
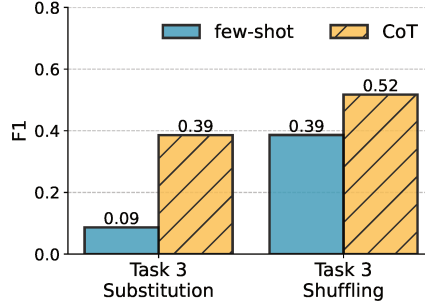
Figure 6: Performance (Accuracy of both) on Tasks 3 when two conflict creation strategies, named entity substitution and entity shuffling, are separately employed, following Figure 4.
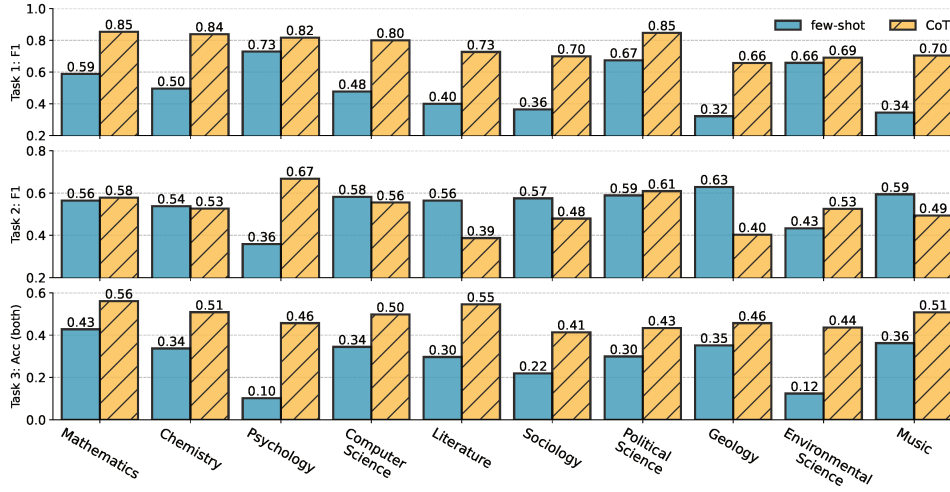

Figure 7: Performance of few-shot and Chain-of-Thought prompting divided into the other ten knowledge domains, following Figure 3.

C Experiment Details

Regarding LLMs' responses, we selectively consider only those that not only provide the correct answer but also adhere to the specified response format as true. For Task 3, we take the named entity substituted or the entity shuffled in the given counter-parametric context as the ground truth for counter-parametric-based answers. Regarding ground truth parametric-based answers, we rely on the responses generated by the LLM when only the question is given instead of utilizing the named entity substituted or the entity shuffled in the corresponding parametric context, which helps eliminate cases where multiple answers are possible for the given question based on parametric knowledge. Instances where the ground truth for parametric-based answers coincides with the ground truth for counter-parametric-based answers are removed.

D Additional Analysis

Conflict Generation Method Breakdown (cont.) In Figure 6, we provide a detailed presentation of the results for Task 3, categorized according to the conflict generation methods. These segmented outcomes for Task 3, consistent with the findings for Task 1 and Task 2, reveal that LLMs demonstrate improved performance in situations where conflicts are introduced through entity shuffling.

Prompt	Few-shot	CoT
...based on the given context and your own knowledge respectively.	0.278	0.495
...considering the provided context and your personal understanding or expertise respectively.	0.248	0.433
Produce two responses...one derived from ... one from ...	0.420	0.483
Produce two responses...one solely based on...one solely based on...	0.455	0.540
Standard Deviations	**0.103**	**0.060**

Table 8: Model performance (Accuracy of both) on Task 3 when different instruction texts are employed.

Domain Breakdown (cont.)   In Figure 7, we present the experimental results across the remaining ten knowledge domains on the three tasks. We can see that the ability to identify knowledge conflicts, accurately locate conflicting information, and generate distinct responses based on conflicting data exhibits variability across the various knowledge domains.

Prompt Consistency (cont.)   Furthermore, we carry out a thorough evaluation of prompt consistency in Task 3 by introducing three additional variations of the instruction, as detailed in Table 8. Overall, the performance remains consistent across these diverse prompts, and CoT exhibits a heightened level of prompt robustness.

Qualitative Analysis   We take a closer look at the experimental results of a particular example under three settings in this section. In this example, the parametric knowledge is "***Pareto optimality*** *is a concept in economics that refers to a state where no individual can be made better off without making someone else worse off. It represents an efficient allocation of resources, where it is impossible to make any further improvements without negatively impacting someone else's welfare.*" and the conflicting knowledge is "***Trade balance*** *is a concept in economics that refers to a state where no individual can be made better off without making someone else worse off. It represents an efficient allocation of resources, where it is impossible to make any further improvements without negatively impacting someone else's welfare.*" In Table 9 and Table 10, the conflicting knowledge is given in the prompt and the expected response is "*Yes*". In Table 11, we provide the conflicting knowledge along with the question "*What is an economic state where no individual can be made better off without making someone else worse off?*" and the expected response is "*Yes*". The same conflicting knowledge and question are also given in Task 3 (Table 12) and the ground truth answers are "*Trade balance*" and "*Pareto efficiency*". We can see that our proposed approach significantly enhances performance.

E Prompt Text

We present all adopted prompt texts in Table 13 to 19 to facilitate reproducibility.

F Multiple Parametric Knowledge Answers

We delve deeper into the possibility of multiple parametric knowledge responses, relaxing the assumption of a singular parametric knowledge answer in the main experiments, and examine how the results may be altered. We sample 600 questions (with conflicts) across different knowledge domains, and employ GPT-3.5-TURBO to generate 3 answers in a zero-shot manner with a temperature of 1.5 to try to induce multiple parametric knowledge answers, following self-consistency (Wang et al., 2023b). We present the results in Table 20. The results show that 25% of the additional parametric answers matched the top answer exactly with 65% of synonyms, while others tend to be invalid or incorrect answers. Therefore, we conclude that the vast majority of questions have only one parametric answer, thus the results and findings won't change significantly if we relax this assumption.

G Multi-hop Questions

Given the prevalence of multi-hop questions in practical scenarios and the difficulty of pinpointing the step where knowledge conflicts occur, we extend our investigation to assess the ability of identifying knowledge conflicts using a two-hop dataset, *Compositional Celebrities* (Press et al., 2022). We randomly choose 400 examples and replace LLMs' responses to one of the sub-questions with a different answer from the in-context corpus, creating 400 positive examples and 400 negative examples. Employing the CoT prompting method (Wei et al., 2022) results in a precision of 0.795, a recall of 0.995, and an F1 score of 0.884 in Task 1 as shown in Table 21. This aligns with our conclusion that LLMs can perform well above random in identifying the existence of knowledge conflicts within contextual information, even under multi-hop question settings.

H Subjective Datasets

We conduct an additional experiment on varying perspectives utilizing the PrimeValue dataset (Sorensen et al., 2024). We randomly sample 350 controversial situations from the dataset including "giving money to your son", "accusing a murderer", "giving blood", "protesting for black lives matter", and "electing Barrack Obama". We elicit parametric viewpoints of GPT-4-turbo and employ GPT-3.5-turbo to create opposite viewpoints following Xie et al. (2023), creating an evaluation dataset of 350 positive samples and 350 negative samples. Employing the CoT prompting method results in a precision of 0.98, a recall of 0.78, and an F1 score of 0.87 in Task 1. Experimental results in Table 22 align with our conclusion that LLMs can perform well above random in identifying the existence of knowledge conflicts, though under this setting, the results are limited to the assumption that the LLM only has one unique firm viewpoint on each situation.

Method	RESPONSE
ZERO-SHOT	No
FEW-SHOT	No
COT	Based on my knowledge, Trade balance is a concept in economics that refers to a state where no individual can be made better off without making someone else worse off; Trade balance represents an efficient allocation of resources, where it is impossible to make any further improvements without negatively impacting someone else's welfare. The number of inconsistency is zero, so the answer is No.
COT + SC	Based on my knowledge, Trade balance is a concept in economics that refers to a state where no individual can be made better off without making someone else worse off; Trade balance represents an efficient allocation of resources, where it is impossible to make any further improvements without negatively impacting someone else's welfare. The number of inconsistency is zero, so the answer is No.
	Based on my knowledge, trade balance is a concept in economics that refers to a state where no individual can be made better off without making someone else worse off; trade balance represents an efficient allocation of resources, where it is impossible to make any further improvements without negatively impacting someone else's welfare. The number of inconsistency is zero, so the answer is No.
	Based on my knowledge, Trade balance is a concept in economics that refers to the difference between the value of a country's exports and the value of its imports; Trade balance represents the state of a country's international trade; Trade balance does not necessarily imply that no individual can be made better off without making someone else worse off; Trade balance does not necessarily represent an efficient allocation of resources. The number of inconsistency is two, which is greater than zero, so the answer is Yes.
	Based on my knowledge, Trade balance is a concept in economics that refers to a state where no individual can be made better off without making someone else worse off. It represents an efficient allocation of resources, where it is impossible to make any further improvements without negatively impacting someone else's welfare. The number of inconsistency is zero, so the answer is No.
	Based on my knowledge, Trade balance is a concept in economics that refers to a state where no individual can be made better off without making someone else worse off; Trade balance represents an efficient allocation of resources, where it is impossible to make any further improvements without negatively impacting someone else's welfare. The number of inconsistency is zero, so the answer is No.
GKP	No
	No
	No
	No
	No

Table 9: Qualitative analysis on Task 1 (Zero-shot, Few-shot, CoT, CoT + SC, and GKP).

Method	RESPONSE
SELF-ASK	What does trade balance refer to? Intermediate answer: Trade balance refers to the difference between a country's exports and imports of goods and services. It is a measure of the net value of a country's trade with the rest of the world. Follow up: Do these two context directly contradict each other? Intermediate answer: No. The term "trade balance" is used in both contexts, but they refer to different concepts. In the given context, trade balance refers to an economic concept, while in my knowledge, it refers to a measure of a country's net trade. So the final answer is: No
BREAK-DOWN	There are two sentences in the given context. Consider the first sentence: Based on what I know, trade balance refers to a state where no individual can be made better off without making someone else worse off. Therefore, there is no conflict in the first sentence. Consider the second sentence: Based on what I know, trade balance represents an efficient allocation of resources, where it is impossible to make any further improvements without negatively impacting someone else's welfare. Therefore, there is no conflict in the second sentence. After considering both sentences, there is no conflict. So the final answer is No.
OURS	generated knowledge: Trade balance is a term used in economics to describe the difference between the value of a country's exports and the value of its imports. A positive trade balance occurs when a country's exports exceed its imports, resulting in a trade surplus. A negative trade balance occurs when a country's imports exceed its exports, resulting in a trade deficit. The trade balance is an important indicator of a country's economic health and can impact its currency value and overall economic stability. step 2: sentence1: Since the generated knowledge includes "Trade balance is a term used in economics to describe the difference between the value of a country's exports and the value of its imports", it can be used to verify the sentence "Trade balance is a concept in economics that refers to a state where no individual can be made better off without making someone else worse off." step 3: sentence1: Sentence1 states that trade balance refers to a state where no individual can be made better off without making someone else worse off, and the generated knowledge states that trade balance refers to the difference between the value of a country's exports and the value of its imports, so there is a conflict, yes. step 4: remaining sentences: "It represents an efficient allocation of resources, where it is impossible to make any further improvements without negatively impacting someone else's welfare." Based on what I know, trade balance represents the difference between the value of a country's exports and the value of its imports, and it does not necessarily represent an efficient allocation of resources or the impossibility of making further improvements without negatively impacting someone else's welfare, so there is a conflict, yes.

Table 10: Qualitative analysis on Task 1 (Self-ask, Break-down, and Ours).

Method	RESPONSE
ZERO-SHOT	No
FEW-SHOT	Yes
CoT	Based on the given context, the answer is trade balance, and based on my knowledge, the answer is also trade balance. So there is no conflict, No.
CoT + SC	Based on the given context, the answer is trade balance, and based on my knowledge, the answer is also trade balance. So there is no conflict, No. Based on the given context, the answer is trade balance, and based on my knowledge, the answer is also trade balance. So there is no conflict, No. Based on the given context, the answer is trade balance, and based on my knowledge, the answer is also trade balance. So there is no conflict, No. Based on the given context, the answer is trade balance, and based on my knowledge, the answer is also trade balance. So there is no conflict, No. Based on the given context, the answer is trade balance, and based on my knowledge, the answer is also trade balance. So there is no conflict, No.
GKP	No No No No No
SELF-ASK	What is the answer to the given question based on the given context? Intermediate answer: Based on the given context, the answer is trade balance. Follow up: What is the answer to the given question based on what you know? Intermediate answer: Based on what I know, the answer is Pareto efficiency. Follow up: Do these two answers directly contradict each other? Intermediate answer: No So the final answer is: No
BREAK-DOWN	There are two sentences in the given context. Consider the first sentence: The first sentence does contain the answer to the given question, which is trade balance. Based on what I know, the answer is also trade balance. Therefore, there is no conflict. There is no need to consider further as I already found the sentence containing the answer to the given question. So the final answer is No.

OURS	disregard the input, solely based on what I know, the answer is Pareto efficiency
	step 2: Since the given question is asking about an economic state where no individual can be made better off without making someone else worse off, the sentence "Trade balance is a concept in economics that refers to a state where no individual can be made better off without making someone else worse off." can be used to answer the given question.
	- I find sentences that can be used to answer the given question
	    step 3: solely based on the sentence "Trade balance is a concept in economics that refers to a state where no individual can be made better off without making someone else worse off.", the answer is trade balance
	    step 4: Since the answer in step 1 is Pareto efficiency and the answer in step 3 is trade balance, there is a conflict, yes.

Table 11: Qualitative analysis on Task 2.

Method	RESPONSE
ZERO-SHOT	The economic state described in the question is known as a trade balance.
FEW-SHOT	Trade balance; Pareto efficiency
CoT	Based on the given context, the answer is trade balance; Based on my own knowledge, the answer is Pareto efficiency.
GKP	Based on the given context, the answer is trade balance; Based on my own knowledge, the answer is Pareto efficiency.
SELF-ASK	What is the answer to the given question based on the given context? Intermediate answer: Based on the given context, the answer is trade balance. Follow up: What is the answer to the given question based on your own knowledge? Intermediate answer: Based on my own knowledge, the answer is Pareto efficiency. So the final answer is: Based on the given context, the answer is trade balance; Based on my own knowledge, the answer is Pareto efficiency.
OURS	In the given context, trade balance is an economic state where no individual can be made better off without making someone else worse off, so ANSWER1 is trade balance. Now disregard the given context and answer the given question "What is an economic state where no individual can be made better off without making someone else worse off?" solely based on your own knowledge. Answer: Solely based on my own knowledge, the economic state where no individual can be made better off without making someone else worse off is called Pareto efficiency, so ANSWER2 is Pareto efficiency.

Table 12: Qualitative analysis on Task 3.

Method	PROMPT
ZERO-SHOT	{context} Does the information above conflict with what you know? Yes/No
Few-shot	Does the given context conflict with what you know? Yes/No Examples: Context: Acetone is an organic compound and the simplest amide derived from acetic acid. It is a white crystalline solid that is soluble in water and commonly used as a solvent in laboratories. It is also utilized in the production of pharmaceuticals, dyes, and plastics. Answer: Yes Context: {context} Answer:
CoT	Does the given context conflict with what you know? Yes/No Examples: Context: Acetone is an organic compound and the simplest amide derived from acetic acid. It is a white crystalline solid that is soluble in water and commonly used as a solvent in laboratories. It is also utilized in the production of pharmaceuticals, dyes, and plastics. Answer: Based on my knowledge, Acetone is an organic compound; Acetone is a type of ketone instead of amide; Acetone is a colorless liquid instead of white solid; Acetone is soluble in water; Acetone is commonly used as a solvent in laboratories; Acetone is utilized in the production of acrylic plastics, polycarbonates and epoxy resins instead of dyes. The number of inconsistency is three, which is greater than zero, so the answer is Yes. Context: {context} Answer:
GKP (KNOWLEDGE GENERATION)	Generate some knowledge about the main entity in the input. Examples: Input: Acetamide is an organic compound and the simplest amide derived from acetic acid. It is a white crystalline solid that is soluble in water and commonly used as a solvent in laboratories. It is also utilized in the production of pharmaceuticals, dyes, and plastics. Knowledge: Acetamide is a organic compound with the chemical formula CH_3CONH_2. It is a white, crystalline solid and is the simplest amide derived from acetic acid. Acetamide is commonly used in laboratory settings as a solvent, and it also has applications in various industries. It can be used in the production of pharmaceuticals, plastics, and as a precursor to other organic compounds. Input: {context} Knowledge:
GKP (KNOWLEDGE INTEGRATION)	Does the given context conflict with what you know? (The given knowledge from you may help with the judgment) Yes/No Examples: Context: Acetamide is an organic compound and the simplest amide derived from acetic acid. It is a white crystalline solid that is soluble in water and commonly used as a solvent in laboratories. It is also utilized in the production of pharmaceuticals, dyes, and plastics. Knowledge: Acetamide is a organic compound with the chemical formula CH_3CONH_2. It is a white, crystalline solid and is the simplest amide derived from acetic acid. Acetamide is commonly used in laboratory settings as a solvent, and it also has applications in various industries. It can be used in the production of pharmaceuticals, plastics, and as a precursor to other organic compounds. Answer: No Context: {context} Knowledge: {knowledge} Answer:
SELF-ASK	Does the given context conflict with what you know? Yes/No Examples: Context: Acetone is an organic compound and the simplest amide derived from acetic acid. It is a white crystalline solid that is soluble in water and commonly used as a solvent in laboratories. It is also utilized in the production of pharmaceuticals, dyes, and plastics. Follow up: What is acetone? Intermediate answer: Acetone is a colorless, flammable liquid solvent widely used in industry and households. It's known for its strong, sweet odor and high volatility. Acetone is utilized as a solvent in paints, varnishes, nail polish removers, and as a key ingredient in the production of plastics, fibers, and pharmaceuticals. Follow up: Do these two context directly contradict each other? Intermediate answer: Yes. In the given context, acetone is a white crystalline solid, while it is a colorless liquid based on what I know. So the final answer is: Yes Context: {context} Follow up:

Table 13: Prompts for Task 1. The number of exemplars in the table may be less than real.

Method	PROMPT
BREAK-DOWN	Does any of the sentence in the given context conflict with what you know? Yes/No Examples: Context: Acetamide is an organic compound and the simplest amide derived from acetic acid. It is a white crystalline solid that is soluble in water and commonly used as a solvent in laboratories. It is also utilized in the production of pharmaceuticals, dyes, and plastics. Answer: There are three sentences in the given context. Consider the first sentence: Based on what I know, acetamide is an organic compound and the simplest amide derived from acetic acid. Therefore, there is no conflict in the first sentence. Consider the second sentence: Based on what I know, acetamide is a white crystalline solid that is soluble in water and commonly used as a solvent in laboratories. Therefore, there is no conflict in the second sentence. Consider the third sentence: Based on what I know, acetamide is utilized in the production of pharmaceuticals, dyes, and plastics. Therefore, there is no conflict in the third sentence. After considering all three sentences, there is no conflict. So the final answer is No. Context: {context} Answer:
OURS	Given the input: step 1: generate some knowledge about the main entity in the input. step 2: from the input, find sentences that can be verified by the generated knowledge step 3: for each sentence found in step2, does it conflict with the generated knowledge? Yes/No step 4: for the remaining sentences that cannot be verified by the generated knowledge, does it conflict with what you know? Input: Acetone is an organic compound and the simplest amide derived from acetic acid. It is a white crystalline solid that is soluble in water and commonly used as a solvent in laboratories. It is also utilized in the production of pharmaceuticals, dyes, and plastics. step 1: generated knowledge: Acetone is a colorless, flammable liquid widely used in industry and households. It's known for its strong, sweet odor and high volatility. Acetone is utilized as a solvent in paints, varnishes, nail polish removers, and as a key ingredient in the production of plastics, fibers, and pharmaceuticals. step 2: sentence1: Since the generated knowledge includes "Acetone is a colorless, flammable liquid widely used in industry and households.", it can be used to verify the sentence "It is a white crystalline solid that is soluble in water and commonly used as a solvent in laboratories." sentence2: Since the generated knowledge includes "Acetone is utilized as a solvent in paints, varnishes, nail polish removers, and as a key ingredient in the production of plastics, fibers, and pharmaceuticals.", it can be used to verify the sentence "It is also utilized in the production of pharmaceuticals, dyes, and plastics." step 3: sentence1: Sentence1 states that acetone is a while crystalline solid, and the generated knowledge states that acetone is a colorless liquid, so there is a conflict, yes. sentence2: Sentence2 states that acetone can be utilized in the production of pharmaceuticals, dyes, and plastics, and the generated knowledge states that acetone is utilized as a solvent in paints and as a key ingredient in the production of plastics and pharmaceuticals, so there is no conflict, no. step 4: remaining sentences: "Acetone is an organic compound and the simplest amide derived from acetic acid." Based on what I know, acetone is an organic compound, but it is not an amide, so there is a conflict, yes. Input: {context} step 1:

Table 14: Prompts for Task 1. The number of exemplars in the table may be less than real.

Method	PROMPT
ZERO-SHOT	{context} Does the information above conflict with what you know regarding the answer to the question {question}? Yes/No
FEW-SHOT	Does the given context conflict with what you know regarding the answer to the question? Yes/No Examples: Context: Acetone is an organic compound and the simplest amide derived from acetic acid. It is a white crystalline solid that is soluble in water and commonly used as a solvent in laboratories. It is also utilized in the production of pharmaceuticals, dyes, and plastics. Question: What is the chemical name of the simplest amide derived from acetic acid? Answer: Yes Context: {context} Question: {question} Answer:
CoT	Does the given context conflict with what you know regarding the answer to the question? Yes/No Examples: Context: Acetone is an organic compound and the simplest amide derived from acetic acid. It is a white crystalline solid that is soluble in water and commonly used as a solvent in laboratories. It is also utilized in the production of pharmaceuticals, dyes, and plastics. Question: What is the chemical name of the simplest amide derived from acetic acid? Answer: Based on the given context, the answer is acetone, while based on my knowledge, the answer is acetamide. So there is a conflict, Yes. Context: {context} Question: {question} Answer:

Table 15: Prompts for Task 2 (Zero-shot, Few-shot, and CoT). At test time, the placeholder {context} is replaced with either parametric knowledge or conflicting knowledge, and the placeholder {question} is replaced by either the question concerning conflicting segments or the question regarding non-conflicting segments. The number of exemplars in the table may be less than real.

Method	PROMPT
GKP (KNOWLEDGE GENERATION)	Generate some knowledge about the input. Examples: Input: What is the chemical name of the simplest amide derived from acetic acid? Knowledge: Acetamide is the chemical name of the simplest amide derived from acetic acid. Input: {question} Knowledge:
GKP (KNOWLEDGE INTEGRATION)	Does the given context conflict with what you know regarding the answer to the given question? (The given knowledge from you may help with the judgment) Examples: Context: Acetone is an organic compound and the simplest amide derived from acetic acid. It is a white crystalline solid that is soluble in water and commonly used as a solvent in laboratories. It is also utilized in the production of pharmaceuticals, dyes, and plastics. Question: What is the chemical name of the simplest amide derived from acetic acid? Knowledge: Acetamide is the chemical name of the simplest amide derived from acetic acid. Answer: Yes Context: {context} Question: {question} Knowledge: {knowledge} Answer:
SELF-ASK	Does the given context conflict with what you know regarding the answer to the given question? Yes/No Examples: Context: Acetone is an organic compound and the simplest amide derived from acetic acid. It is a white crystalline solid that is soluble in water and commonly used as a solvent in laboratories. It is also utilized in the production of pharmaceuticals, dyes, and plastics. Question: What is the chemical name of the simplest amide derived from acetic acid? Follow up: What is the answer to the given question based on the given context? Intermediate answer: Based on the given context, the answer is acetone. Follow up: What is the answer to the given question based on what you know? Intermediate answer: Based on what I know, the answer is acetamide. Follow up: Do these two answers directly contradict each other? Intermediate answer: Yes So the final answer is: Yes Context: {context} Question: {question} Follow up:

Table 16: Prompts for Task 2 (GKP and Self-ask). At test time, the placeholder {context} is replaced with either parametric knowledge or conflicting knowledge, {question} is replaced by either the question concerning conflicting segments or the question regarding non-conflicting segments, and {knowledge} is substituted with the knowledge produced during the GKP knowledge generation step. The number of exemplars in the table may be less than real.

Method	PROMPT
BREAK-DOWN	Does any of the sentence in the given context conflict with what you know regarding the answer to the given question? Yes/No Examples: Context: Acetone is an organic compound and the simplest amide derived from acetic acid. It is a white crystalline solid that is soluble in water and commonly used as a solvent in laboratories. It is also utilized in the production of pharmaceuticals, dyes, and plastics. Question: What is the chemical name of the simplest amide derived from acetic acid? Answer: There are three sentences in the given context. Consider the first sentence: The first sentence does contain the answer to the given question, which is acetone. Based on what I know, the answer is acetamide. Therefore, there is a conflict. There is no need to consider further as I already found the sentence containing the answer to the given question. So the final answer is Yes. Context: {context} Question: {question} Answer:
OURS	Given the input and the question: step 1: disregard the input, answer the given question solely based on what you know step 2: from the input, find sentences that can be used to answer the given question if any - if you find sentences that can be used to answer the given question: step 3: extract the answer from these sentences step 4: does the answer in step 1 conflict with the answer in step 3? Yes/No - if you don't find sentences that can be used to answer the given question: step 3: return there is no conflict, no. Input: Acetone is an organic compound and the simplest amide derived from acetic acid. It is a white crystalline solid that is soluble in water and commonly used as a solvent in laboratories. It is also utilized in the production of pharmaceuticals, dyes, and plastics. Question: What is the chemical name of the simplest amide derived from acetic acid? step 1: disregard the input, solely based on what I know, the answer is acetamide step 2: Since the given question is asking about the chemical name of the simplest amide derived from acetic acid, the sentence "Acetone is an organic compound and the simplest amide derived from acetic acid." can be used to answer the given question. - I find sentences that can be used to answer the given question step 3: solely based on the sentence "Acetone is an organic compound and the simplest amide derived from acetic acid.", the answer is acetone step 4: Since the answer in step 1 is acetamide, and the answer in step 3 is acetone, there is a conflict, yes. Input: {context} Question: {question} step 1:

Table 17: Prompts for Task 2 (Break-down and Ours). The placeholder {context} is substituted for either parametric knowledge or conflicting knowledge, while the placeholder {question} is substituted for either the question about the conflicting segments or the question about the non-conflicting segments at the test time. The number of exemplars in the table may be less than real.

Method	PROMPT
ZERO-SHOT	Context: {context} Question: {question} Answer the given question based on the given context and your own knowledge respectively. Answer:
FEW-SHOT	Answer the question based on the given context and your own knowledge respectively. Examples: Context: Acetone is an organic compound and the simplest amide derived from acetic acid. It is a white crystalline solid that is soluble in water and commonly used as a solvent in laboratories. It is also utilized in the production of pharmaceuticals, dyes, and plastics. Question: What is the chemical name of the simplest amide derived from acetic acid? Answer: acetone; acetamide Context: {context} Question: {question} Answer:
COT	Answer the question based on the given context and your own knowledge respectively. Examples: Context: Acetone is an organic compound and the simplest amide derived from acetic acid. It is a white crystalline solid that is soluble in water and commonly used as a solvent in laboratories. It is also utilized in the production of pharmaceuticals, dyes, and plastics. Question: What is the chemical name of the simplest amide derived from acetic acid? Answer: Based on the given context, the answer is acetone; Based on my own knowledge, the answer is acetamide. Context: {context} Question: {question} Answer:
GKP	Answer the question based on the given context and your own knowledge respectively. (You may find the given knowledge from you helpful.) Examples: Context: Acetone is an organic compound and the simplest amide derived from acetic acid. It is a white crystalline solid that is soluble in water and commonly used as a solvent in laboratories. It is also utilized in the production of pharmaceuticals, dyes, and plastics. Question: What is the chemical name of the simplest amide derived from acetic acid? Knowledge: Acetamide is the chemical name of the simplest amide derived from acetic acid. Answer: Based on the given context, the answer is acetone; Based on my own knowledge, the answer is acetamide. Context: {context} Question: {question} Knowledge: {knowledge} Answer:

Table 18: Prompts for Task 3 (Zero-shot, Few-shot, CoT, and GKP). During the test phase, the placeholder {context} is replaced with either parametric knowledge or conflicting knowledge, while the placeholder {question} is replaced with either the question pertaining to conflicting segments or the question regarding non-conflicting segments. Similarly, {knowledge} is replaced with the knowledge that is generated, as in Task 2. The number of exemplars in the table may be less than real.

Method	PROMPT
SELF-ASK	Answer the question based on the given context and your own knowledge respectively. Examples: Context: Acetone is an organic compound and the simplest amide derived from acetic acid. It is a white crystalline solid that is soluble in water and commonly used as a solvent in laboratories. It is also utilized in the production of pharmaceuticals, dyes, and plastics. Question: What is the chemical name of the simplest amide derived from acetic acid? Follow up: What is the answer to the given question based on the given context? Intermediate answer: Based on the given context, the answer is acetone. Follow up: What is the answer to the given question based on your own knowledge? Intermediate answer: Based on my own knowledge, the answer is acetamide. So the final answer is: Based on the given context, the answer is acetone; Based on my own knowledge, the answer is acetamide. Context: {context} Question: {question} Follow up:
OURS	Generate two answers to the given question: ANSWER1 solely based on the given context and ANSWER2 solely based on your own knowledge. Examples: Question: What is the chemical name of the simplest amide derived from acetic acid? Answer the question solely based on the given context: Acetone is an organic compound and the simplest amide derived from acetic acid. It is a white crystalline solid that is soluble in water and commonly used as a solvent in laboratories. It is also utilized in the production of pharmaceuticals, dyes, and plastics. Answer: In the given context, acetone is the simplest amide derived from acetic acid, so ANSWER1 is acetone. Now disregard the given context and answer the given question "What is the chemical name of the simplest amide derived from acetic acid?" solely based on your own knowledge. Answer: Solely based on my own knowledge, the chemical name of the simplest amide derived from acetic acid is acetamide, so ANSWER2 is acetamide. Question: {question} Answer the question solely based on the given context: {context} Answer:

Table 19: Prompts for Task 3 (Self-ask and Ours). During the test phase, the placeholder {context} is replaced with either parametric knowledge or conflicting knowledge, while the placeholder {question} is replaced with either the question pertaining to conflicting segments or the question regarding non-conflicting segments. The number of exemplars in the table may be less than real.

Domain	ExactMatch	Synonyms	Invalid/Incorrect
Mathematics	8	17	5
Biology	9	18	3
Chemistry	7	20	3
Physics	4	23	3
Psychology	6	20	4
Computer Science	7	21	2
History	11	18	1
Literature	15	14	1
Sociology	3	27	0
Economics	9	19	2
Political Science	10	16	4
Philosophy	5	21	4
Anthropology	6	24	0
Geology	8	17	5
Linguistics	1	25	4
Art History	12	10	8
Environmental Science	8	19	3
Health Sciences	7	22	1
Communications	9	20	1
Music	3	20	7
Total	**148**	**391**	**61**

Table 20: Consistency of 3 parametric knowledge answers generated by GPT-3.5-TURBO on 600 randomly sampled questions. The majority of additional answers are found to be nearly the same, showing that, without the assumption of unique parametric knowledge answer, our conclusions won't change significantly.

Method	n	TP	TN	FP	FN	Acc	Precision	Recall	F1
ZERO-SHOT	743	147	298	73	225	0.599	0.668	0.395	0.497
FEW-SHOT	796	164	224	174	234	0.487	0.485	0.412	0.446
COT	756	373	285	96	2	0.870	0.795	0.995	0.884

Table 21: Results of the Compositional Celebrities dataset (Press et al., 2022) on Task 1, which aligns with our conclusion that LLMs exhibit proficiency beyond random chance in detecting the presence of knowledge conflicts.

Method	n	TP	TN	FP	FN	Acc	Precision	Recall	F1
ZERO-SHOT	698	61	348	1	288	0.59	0.98	0.18	0.30
FEW-SHOT	700	173	347	3	177	0.74	0.98	0.49	0.66
COT	700	272	345	5	78	0.88	0.98	0.78	0.87

Table 22: Results of the PrimeValue dataset (Sorensen et al., 2024) on Task 1, which aligns with our conclusion that LLMs can perform well above random in identifying the existence of knowledge conflicts.