

Understanding the Interplay between Parametric and Contextual Knowledge for Large Language Models

Sitao Cheng[♡], Liangming Pan^{♠*}, Xunjian Yin[♡], Xinyi Wang[♡], William Yang Wang[♡]

[♡]University of California, Santa Barbara [♠]University of Arizona

{sitaocheng, yinxunjian, xinyi_wang, william}@ucsb.edu

liangmingpan@arizona.edu

Abstract

Large language models (LLMs) encode vast amounts of knowledge during pre-training (parametric knowledge, or PK) and can further be enhanced by incorporating contextual knowledge (CK). Can LLMs effectively integrate their internal PK with external CK to solve complex problems? In this paper, we investigate the dynamic interaction between PK and CK, categorizing their relationships into four types: *Supportive*, *Complementary*, *Conflicting*, and *Irrelevant*. To support this investigation, we introduce ECHOQA, a benchmark spanning scientific, factual, and commonsense knowledge. Our results show that LLMs tend to suppress their PK consistently when contextual information is available, even when it is complementary or irrelevant. While tailored instructions can encourage LLMs to rely more on their PK, they still struggle to fully leverage it. These findings reveal a key vulnerability in LLMs, raising concerns about their reliability in knowledge-intensive tasks. We will release our code and dataset to facilitate future research.¹

1 Introduction

Large language models (LLMs) (Ouyang et al., 2022; Touvron et al., 2023; OpenAI, 2024) capture a substantial amount of knowledge through pre-training on large-scale corpus, referred to as *parametric knowledge* (PK). When applied to problem-solving, LLMs often rely on additional information provided as context, known as *contextual knowledge* (CK)². A critical area that remains underexplored is how effectively LLMs can integrate PK and CK to solve complex problems. This ability is crucial for human-like intelligence, where we effortlessly draw upon stored knowledge to adapt to new environments, reflecting our capacity to

perceive, recall, and integrate information when encountering new stimuli (Gibson, 1988). Effective collaboration between PK and CK is also vital for real-world applications, such as retrieval-augmented generation (RAG) (Schick et al., 2023; Trivedi et al., 2023; An et al., 2024).

However, we find that for LLMs, an effective interaction between PK and CK can be challenging. For example, we observe a counter-intuitive phenomenon that when supplementing an LLM with more task-relevant knowledge in the context (CK), sometimes the performance degrades because the LLM fails to effectively utilize the parametric knowledge (PK). Examples are shown in Figure 1. This catastrophic test time forgetting greatly limits LLMs capability to effectively solve user queries when relevant context is provided. In this paper, our objective is to thoroughly investigate *the intricate dynamics between PK and CK*, leading to a better understanding of how well LLMs integrate and prioritize these sources of knowledge. Previous works separately study when CK is conflicting or irrelevant to PK, raising concerns about the vulnerability (Xie et al., 2024; Wu et al., 2024). However, they fail to answer more fine-grained questions such as: To what extent are LLMs aware of PK, in the presence of CK? How well can LLMs leverage PK given CK? What factors can affect such ability?

To perform a systematic study, we formulate *reasoning types* based on relationships between CK and PK-*Supportive*, *Complementary*, *Conflicting* and *Irrelevant*, illustrated by examples in Figure 1 and Table 1. Due to the lack of appropriate datasets to examine this ability, we introduce a new benchmark **ECHOQA** (Section 3.3) spanning diverse knowledge-intensive reasoning tasks, *i.e.*, scientific, factual and commonsense, across various LLMs. We extract PK by ensuring LLMs can reach 100% performance for the knowledge. Next, we construct CK from PK to fit our reasoning types, as in the middle of Figure 1. We hope that ECHOQA will serve

*Corresponding Author

¹https://github.com/sitaocheng/Knowledge_Interplay

²For simplicity, PK and CK denote parametric and contextual knowledge throughout this paper.



Figure 1: Our benchmark ECHOQA, accessing LLMs ability to echo their parametric knowledge (PK) when contextual knowledge (CK) is present. We firstly question LLMs to obtain PK and discard knowledge they cannot answer. Then, we construct CK by various *reasoning types* (Table 1). Next, we question LLMs given CK. Exemplar result is by Llama 3.1-70B on ALCUNA (Yin et al., 2023).

as a valuable resource to steer future explorations on LLMs leverage and integration of PK with CK.

To investigate to what the extent LLMs can leverage PK, we design a series of reasoning instructions, which progressively enforce the use of PK, guiding LLMs preference to varying degrees, as detailed in Section 3.2. Furthermore, we study other factors affecting LLMs abilities to leverage their internal knowledge, *e.g.*, popularity of entities and methods of contextual knowledge fabrication, shedding lights on future improvements (Section 5). Our investigation draws the following findings:

- *The suppression of PK in the presence of CK is universal*, regardless of the model, knowledge type, instructions or reasoning type (Complementary, Conflicting or Irrelevant). LLMs *disregard* their own knowledge, *e.g.*, more than 60% of cases in scientific knowledge for all tested models, relying solely on the context for reasoning. This highlights the vulnerability of LLMs in leveraging PK.
- *LLMs are more likely to recall some types of knowledge*, *e.g.*, commonsense knowledge. We find evidence indicating that the reason behind is likely the imbalance in training corpus.
- *Explicit instructions can help recall more PK, but still far from fully leveraging PK*. This implies more sophisticated framework design has the potential to solve this problem to a larger extent. We show example prompt templates in Appendix 19.

2 Related Work

Parametric Knowledge in LLMs is encoded in models parameters through vast amounts of pre-train data. Previous research extensively explores how LLMs leverage PK for reasoning (Li et al., 2024; Yang et al., 2024b; Wang et al., 2024). Some reveal the challenges in leveraging PK (Wang et al., 2023a; Allen-Zhu and Li, 2023). These studies

focus on models ability to utilize PK to perform various tasks, *e.g.*, multi-hop reasoning, without considering the influence of contextual knowledge (CK). While valuable, these analyses do not address how PK is dynamically applied in real-world contexts where CK also plays an important role. Moreover, PK embeded in language models can be outdated or incorrect, prone to hallucination (Elazar et al., 2021; Lazaridou et al., 2021; Zhong et al., 2023).

Contextual Knowledge for LLMs refers to the information augmented in LLMs context window to enhance generation (RAG), which hopefully supplements LLMs internal PK to improve faithfulness (Gua et al., 2020; Qin et al., 2023; Mallen et al., 2023). Previous work primarily focuses on optimizing the retrieval of relevant documents for reasoning (Press et al., 2023; Asai et al., 2023; Zhuang et al., 2024). However, as documents contain diverse information, LLMs may be confused when reasoning, even with optimal retrieval, *e.g.*, the given CK contradicts PK (Xu et al., 2024). In this paper, we find that even complementary CK can hinder LLMs ability to fully leverage PK. We formulate the relationship between PK and CK to investigate the influence of different kinds of CK on LLMs utilization of PK.

Interplay between Parametric and Contextual Knowledge mainly includes behavioral analysis or fusion of both knowledge. Yin et al. (2023) assesses models reaction on new knowledge, focusing on knowledge construction. Some elicit PK from LLMs and design frameworks to introduce conflict (Wang et al., 2023b; Xie et al., 2024) or irrelevant CK (Wu et al., 2024). Others explore the fusion of CK and PK for improved sufficiency (Jiang et al., 2023; Zhang et al., 2024). However, LLMs awareness and utilization of PK under various CK is understudied. Besides, the focus on single type

Reasoning Type	Example	Metric
<i>Supportive</i> $PK \odot CK$	Q: What’s the shape of the earth? A: Spherical PK: The earth is spherical. CK: The earth is spherical.	Acc, U_R
<i>Complementary</i> $PK \wedge CK$	Q: The shape of the planet Z-man was born on? A: Spherical PK: The earth is spherical. CK: Z-man was born on earth.	Acc, U_R
<i>Conflicting</i> $PK \oplus CK$	Q: What’s the shape of the earth? A_{ck} : Flat; A_{pk} : Spherical PK: The earth is spherical. CK: The earth is flat.	M_R
<i>Irrelevant</i> $PK \vee CK$	Q: What’s the shape of the earth? A: Spherical PK: The earth is spherical. CK: The earth orbits around the sun.	Acc, U_R

Table 1: Definition, examples and metrics of our designed reasoning types over various relationships between CK and PK. Note that we ensure PK is known by LLMs and not present in context. U_R, M_R, A_{ck}, A_{pk} denotes unknown ratio, memorization ratio, answers faithful to CK and PK, respectively.

of knowledge may overlook the intricate relations between the two knowledge sources reflected real-world scenarios, *e.g.*, LLM-based search engines or chatbots (Nakano et al., 2021) or human reasoning process, *e.g.*, human can naturally integrate knowledge in brain with new environments. We aim to comprehensively examine LLMs ability to manage and integrate the two knowledge sources.

3 Experimental Setup

As shown in Table 1, we design four *reasoning types* based on various relationships between PK and CK. Without the loss of generality, we adopt Question Answering (QA) as the reasoning task, following prior work (Mallen et al., 2023; Cheng et al., 2024). Given a question Q and CK, an LLM leverages PK and CK to reason the answer A . Formally, if Q is posed, then CK operated with PK under reasoning type (rt) leads to A :

$$Q \rightarrow (PK \text{ } rt \text{ } CK \rightarrow A) \quad (1)$$

where $rt \in \{\odot, \wedge, \oplus, \vee\}$ and PK always holds true, meaning that we ensure LLM has PK required for Q for all settings. In experiments, we vary CK based on our designed rt as shown in Figure 1. Different from Xie et al. (2024), we only present CK in the context, to assess the ability to utilize PK naturally.

\odot **Supportive** We have CK and PK convey the same information. Prior experiments show that models behave consistently with or without CK (as expected, or LLMs can effectively adopt PK). Therefore, we do not report this type in main results. Refer to Appendix B.2 for statistics.

\wedge **Complementary** This type requires both PK and CK for (multi-hop) reasoning. We maintain knowledge sufficiency by assuring PK is known by LLM and CK is new to LLMs. A key is to establish meaningful connections between them. Previous study construct multi-hop questions by replacing entities with descriptions (Talmor and Berant, 2018; Huang et al., 2023), which does not ensure that CK is new to LLMs. To address this, we create synthetic entity relating to PK. For the example in Table 1, given PK “the earth is spherical”, the CK “Z-man was born on earth” is complementary to PK to answer “The shape of the planet Z-man was born on”.

\oplus **Conflicting** A fact in CK is fabricated from PK. For the example in Figure 1, we modify “Michael Jordan is a Basketball Player” into “Michael Jordan works as an ESPN basketball commentator for 20 years”. We also create longer narratives around the fabricated fact. We ensure Q require that exact fact to answer. We denote the answer that aligns with CK and PK as A_{ck} and A_{pk} , respectively.

\vee **Irrelevant** We randomly sample irrelevant information as CK, under certain constraints, *e.g.*, CK for questions about *occupation* should still pertain to occupation. As our goal is to test LLMs ability to leverage PK, we only consider the case when PK is necessary for Q in main results. We discuss more about knowledge filtering ability in Appendix B.5.

3.1 Evaluation Metrics

We design various metrics to assess LLMs ability to leverage PK, as shown in Table 1. Specifically: **Accuracy** (Acc) evaluates whether the output ex-

actly matches the ground truth answer.

Memorization Ratio (M_R) measures to what extent the model adheres to PK given conflict CK. It is calculated as $M_R = \frac{f_{pk}}{f_{pk} + f_{ck}}$, where f_{pk} and f_{ck} denotes the frequency of answer faithful to A_{ck} and A_{pk} , respectively, following Longpre et al. (2021). **Unknown Ratio** (U_R) measures the ratio of models generating unknown. U_R reveals models uncertainty to some extent. Moreover, since we make sure PK is required by Q , higher U_R shows lower perception of PK.

3.2 Progressively-Enforced Instructions

Recent post-training techniques, *e.g.*, instruction-tuning (Chung et al., 2024), enable LLMs to fluently follow human instructions. To explore to what extent LLMs can leverage PK, we design a series of progressively-enforced instructions to guide models (detailed templates in Appendix C).

- **Neutral Instruction** is natural and straightforward. We treat LLMs as a helpful assistant with knowledge to answer questions. Note that we indeed ask LLMs to use their own knowledge.
- **Trust Yourself Instruction** proactively shows the insufficiency or asks LLMs to trust their own knowledge. Concerningly, such enforcing instruction is to test LLMs ability, but not for practical use where the sufficiency and trustworthiness of knowledge are not ensured.
- **Speak Out Loud Instruction** further pushes LLMs to firstly explicitly utter their PK then trust themselves. This differs from Xie et al., 2024, who asks models to make a choice when PK and CK are both in context as a behavioral study. Instead, we test LLMs ability to output PK. Also, this enforcing instruction is not for practical use.

3.3 ECHOQA

We introduce our dataset **ECHOQA** to test LLMs ability to echo their PK when encountering CK. Following Xie et al., 2024, questions are in multi-choice format and provided an unknown option. The key step is to firstly obtain models PK by questioning about a knowledge triple, where a correct answer indicates that the LLM possesses the relevant knowledge, as in the left of Figure 1. Then, we construct CK from PK based on our designed reasoning types, as in the middle of Figure 1. We conduct human evaluations on the quality of the constructed CK and Q , showing that 97.6% and 95.3% of Q and CK perfectly fit our reasoning types (Table 1). Notably, our construction can adapt to any

knowledge-intensive task. Specifically, ECHOQA is adapted from ALUNA (Yin et al., 2023), ConflictQA (Xie et al., 2024), MuSiQue (Trivedi et al., 2022), OpenBookQA (Mihaylov et al., 2018), covering scientific, factual and commonsense knowledge. Detailed descriptions are in Appendix A.

3.4 Language Models

We include multiple LLMs spanning diverse parameter scales, *i.e.*, two closed-source (o1 (OpenAI, 2024), GPT-4o and GPT-4o-mini (OpenAI, 2023)) and three open-source (Llama 3.1-70B, Llama 3.1-8B (Meta, 2024) and Qwen 2-7B (Yang et al., 2024a)). These models have demonstrated strong performance in knowledge-intensive tasks and instruction-following capabilities, making them well-suited for our tasks. In ECHOQA, we filter PK for each LLM (as in the left of Figure 1). Statistics of examples are detailed in Table 4. For all LLMs, we experiment with Chain-of-Thought (Wei et al., 2022) for analysis and adopt *zero-shot* setting to avoid the bias introduced by demonstrations. Other setups remain as default. Specific model API/checkpoints are listed in Appendix B.

4 Experimental Results

4.1 Complementary Reasoning

Although provided with complementary context, LLMs still suppress their own knowledge. Figure 2 illustrates the Accuracy for questions requiring both their parametric (PK) and contextual knowledge (CK). Table 2 shows the Unknown Ratio U_R on ALCUNA (more in Appendix B.3, B.4). We also experiment with zero knowledge and all the required knowledge in the context as the lower and upper bound (blue and orange bars), respectively. The following observations can be drawn:

First, *LLMs cannot fully leverage PK, even though PK with the given CK is sufficient for the question.* All tested LLMs across all knowledge types demonstrate an upward trend in Accuracy, with stronger models exhibiting higher absolute performance. With all required knowledge in context, LLMs perform effectively (orange bar). However, the performance drops dramatically with only CK in context (red/green bar). Figure 2 (a) (scientific knowledge) shows that the introduction of complementary CK even degrades the performance for most LLMs, compared to zero knowledge. We scrutinize 200 samples and find that the presence of CK hinders LLMs from engaging in deeper rea-

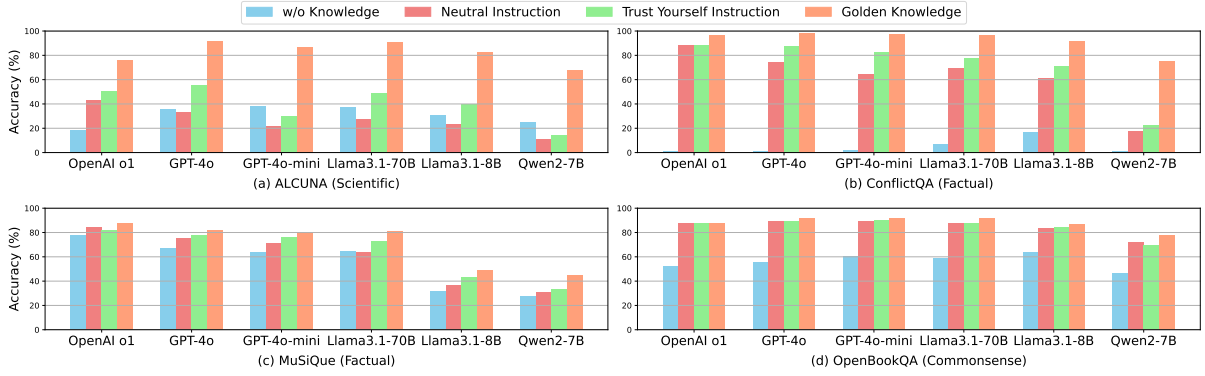


Figure 2: Accuracy for Complementary Reasoning. *w/o* Knowledge and Golden Knowledge means no information and all the required information is given, respectively. The upward trend shows that LLMs suppress PK even with complementary CK, comparing to the orange bar.

soning, *e.g.*, inferring entity properties by names, leading to an over-reliance on context. We further discuss the phenomenon in Section 5.1.

Second, *with complementary CK, LLMs show even increased uncertainty*. Table 2 indicates that the complementary CK “confuses” most tested LLMs significantly, doubling the U_R for Llama and Qwen models, comparing “NI” with “*w/o* K” column. This implies that LLMs depend excessively on the context to seek answers, suppressing their own PK even though PK is required for the question.

Instruction-following aids in recalling some memories, but still way off. By asking LLMs to adopt PK in a progressively-enforced tone, our designed instructions (Section 3.2) significantly bring up the performance, demonstrating the strong instruction-following capabilities of LLMs and showing that instructions can modulate perception of knowledge to some extent. However, even when explicitly guided, LLMs still cannot fully recall the necessary PK they already have, leaving a large margin compared to when all knowledge is provided in the context. This highlights the need for dedicate prompt or framework design for better integration of different sources of knowledge.

Commonsense knowledge is easier to remember, while the scientific and factual is harder. Comparing different types of knowledge in Figure 2, we observe that LLMs exhibit nearly doubled accuracy in commonsense relative to scientific knowledge (red and green bar Figure 2 (a)(d)). We speculate the higher recall of specific PK is likely because such information appears more frequently in the training corpus (more discussions in Section 5.1). Moreover, the upper bound of MuSiQue is considerably lower, especially for weak models, while the upward trend persists, suggesting that multi-step questions mainly challenge reasoning

Models	<i>w/o</i> K	NI	TYI	Gold K
<i>Closed-source LLMs</i>				
OpenAI o1	65.78	46.12	35.43	20.63
GPT-4o	36.45	59.90	26.94	8.08
GPT-4o-mini	46.12	76.13	68.28	12.52
<i>Open-source LLMs</i>				
Llama 3.1-70B	23.89	62.72	23.88	0.08
Llama 3.1-8B	18.12	58.50	34.35	12.70
Qwen 2-7B	40.28	81.26	73.62	28.60

Table 2: Unknown Ratio U_R (%) for Complementary Reasoning on ALCUNA. *w/o* K, Gold K means none and golden knowledge is given, respectively. NI, TYI denotes Neutral and Trust Yourself Instruction, respectively. The **bold** denotes the highest U_R in each row.

ability instead of the knowledge recall.

4.2 Conflicting Reasoning

LLMs rarely trust themselves when faced with conflicting context. Figure 3 depicts the ratio of LLMs relying on their parametric knowledge (PK) when presented with conflicting contextual knowledge (CK). Under Neutral Instruction, other than the strong OpenAI o1, few LLMs lean on themselves for more than 10% cases across all datasets. Note that the unknown ratio here is relatively low (Appendix B.4), making it reasonable to draw conclusions from Figure 3. It is demonstrated that *LLMs are faithful to the conflict context across all types of knowledge*, including commonsense, echoing findings in prior work (Xie et al., 2024). This observation underscores limitations in LLMs confidence and usage on their own knowledge when confronted with disinformation, warranting dedicated content filtering and careful safety design for LLM-based systems.

Instructions significantly influence LLMs preferences, but still struggle to fully recall their

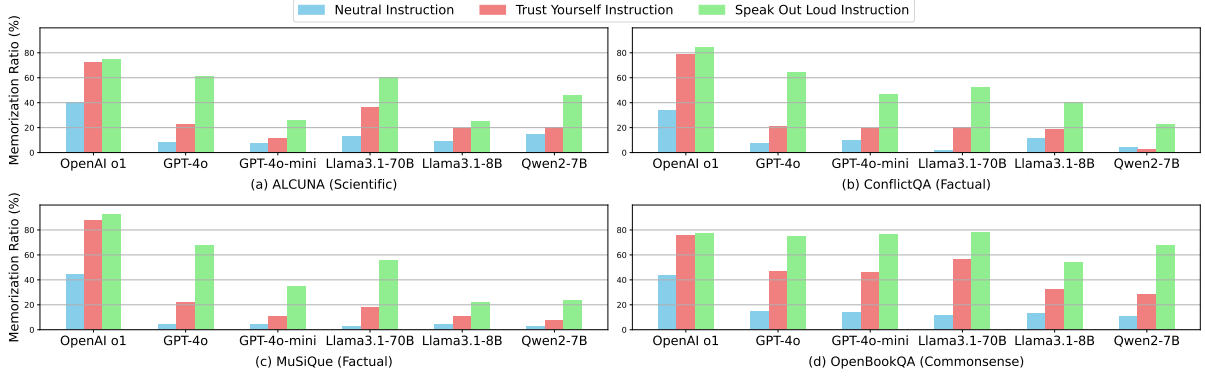


Figure 3: Memorization Ratio for Conflicting Reasoning. LLMs rarely trust themselves (PK) when faced with conflicting CK, though instructions modulate their preference to some extent.

memory. We further explore to what extent LLMs can recall PK when faced with conflict CK. As we instruct LLMs to trust themselves in a progressively more enforced tone, the Memorization Ratio (M_R) consistently goes up, with stronger model exhibiting relatively higher M_R . Compared to Neutral Instruction, the “Trust Yourself Instruction” almost doubles the M_R , with the “Speak Out Loud Instruction” further boosting M_R . However, despite instructed to trust themselves, few of M_R exceed 60% for scientific and factual knowledge, indicating that LLMs still cannot fully recall their internal knowledge when the context is in conflict. Again, o1 exhibits more confidence than others, especially when told to “Trust Yourself”, but there is still space for improvement. This also highlights the vulnerability of LLMs prone to disinformation.

LLMs exhibit greater resistance to conflicts in commonsense knowledge. Figure 3(d) shows that LLMs can recall commonsense knowledge relatively better, with our instructions leading to stronger enhancement. Specifically, most tested LLMs are confident against conflicting CK in over 60% cases with the “Speak Out Loud Instruction” (green bar). Since commonsense facts are less likely to become outdated or change over time, it is crucial that LLMs stand on themselves when encountering conflicting information. Although not perfect, this offers promising insights into preventing attacks. In contrast, while the upward trend still holds, most LLMs are less likely to trust their own factual knowledge, often succumbing to conflicting CK (Figure 3(b)(c)), highlighting the importance of knowledge authenticity in LLM-based systems.

4.3 Irrelevant Reasoning

Although having the knowledge, LLMs still seek answers in the irrelevant context. Figure 4 illustrates the Accuracy for questions requiring LLMs

Models	NI	TYI	SOLI
<i>Closed-source LLMs</i>			
OpenAI o1	6.12	6.12	0.98
GPT-4o	50.38	13.46	0.53
GPT-4o-mini	48.42	17.81	0.80
<i>Open-source LLMs</i>			
Llama3.1-70B	55.14	27.36	2.33
Llama3.1-8B	50.41	23.88	4.83
Qwen2-7B	80.59	41.82	4.57

Table 3: Unknown Ratio U_R (%) for Irrelevant Reasoning on ConflictQA. NI, TYI, SOLI denotes Neutral, Trust Yourself and Speak Out Loud Instruction, respectively. The **bold** denotes highest U_R in each row.

knowledge (PK), when provided with one to four irrelevant documents in context (CK). Table 3 shows the Unknown Ratio (U_R) on ConflictQA (more in Appendix B.4). Note that we indeed ask LLMs to adopt their own knowledge in all instructions. Under “Neutral Instruction”, although holding the answers and the provided context is irrelevant, most LLMs, other than o1, tend to search for answers solely on the CK, particularly in the case of scientific knowledge (blue bar). This indicates *the challenge LLMs face in disentangling PK from CK during reasoning*. Again, commonsense knowledge is more easily recalled, with GPT and Llama achieving up to 80% accuracy on “Neutral Instruction”. We further discuss LLMs filtering ability in Appendix B.5).

We also observe that when the given information is irrelevant, other than o1, LLMs exhibit high uncertainty, with more than 50% Unknown Ratio for most tested models. With our progressively-enforced instructions, the uncertainty decreases both gradually and substantially, to below 5% for all tested models on ConflictQA (Table 3). This also suggests higher possibility of models leverage of PK. Moreover, OpenAI o1 shows extraordinary

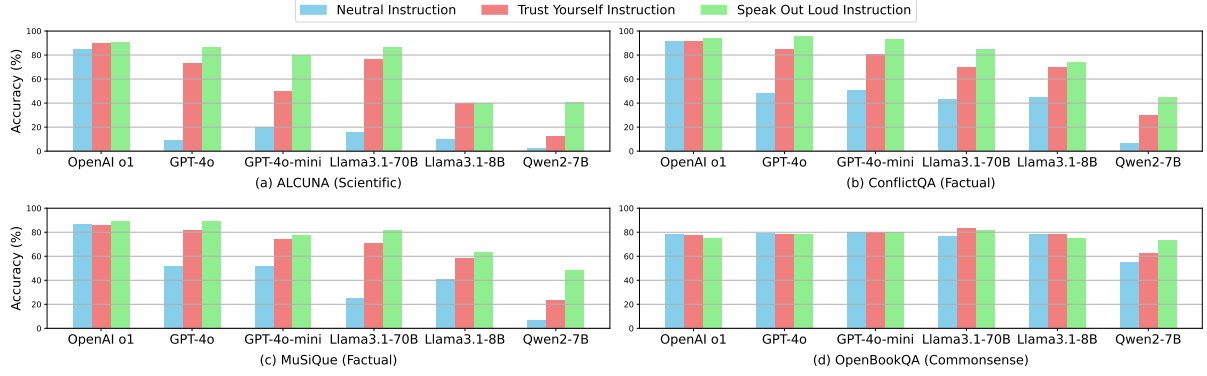


Figure 4: Accuracy for Irrelevant Reasoning, showing that LLMs relying on CK even though it is irrelevant and that instructions can substantially modulate their leverage of knowledge.

ability echoing PK when CK is irrelevant, showing close results at different levels of instructions. This indicates the intricate design behind o1 and sheds lights on future improvements.

LLMs echo their knowledge relatively better when the context is irrelevant. Comparing the absolute performance with Conflicting and Complementary Reasoning (Figure 2 and 3), LLMs are able to better leverage PK for reasoning when CK is irrelevant. With our progressively-enforced instructions, LLMs further improve their memory recall by a large margin (Accuracy reaches 80% for most tested models Figure 4) and become more certain about their answers (Unknown Ratio drops drastically in Table 3). The observation, alongside the results for Complementary and Conflicting Reasoning, indicates the possibility of leveraging PK for reasoning. However, any knowledge presented in context tends to capture their attention to some extent. In other words, *the relationships between PK and CK greatly influence LLMs awareness and leverage of their knowledge.*

5 Discussion

5.1 Factors Affecting Models Leverage of Parametric Knowledge

In previous findings, we show that the awareness and utilization of PK may be influenced by instructions, knowledge type, and the intricate relations with CK. Here, we discuss additional factors:

LLMs can recall popular knowledge more effectively. Inspired by [Mallen et al. \(2023\)](#), LLMs may lean on more popular entities, *i.e.*, monthly associated Wikipedia page views. We evaluate representative LLMs with Conflicting Reasoning on ConflictQA *w.r.t.* popularity of the topic entity in the question in Figure 5. Intuitively, LLMs can recall memories better for more popular questions.

This upward trend holds across all LLMs and our designed instructions. Specifically, GPT-4o-mini and Llama 3.1-70B achieve nearly 80% Memorization Ratio for the most popular questions, almost doubling that of the least popular ones. This observation also indicates that the reason behind the better recall of commonsense knowledge is likely the knowledge frequency in training corpus.

Introduction of relevant knowledge can draw overly excessive attention to the context. As mentioned in Section 4.1 (Figure 2(a)), counter-intuitively, even though PK with the given CK is sufficient to the question, the performance can still degrade compared to direct answering. We manually examine 200 random samples generated by Llama 3.1-70B on ALCUNA where LLM is able to answer on their own but failed when complementary CK is introduced. Note that we do ask LLMs to use their own knowledge in the instruction (Neutral Instruction). We discover that in 93.5% of cases, LLMs relied solely on the context and answer choices and even claimed the insufficiency of knowledge (Appendix B.7). This suggests that *LLMs tend to ignore their knowledge, inhibit deeper thinking and depend excessively on the context.*

We speculate that the phenomenon may stem from post-training where LLMs are fine-tuned in a “*Instruction; Context(optional); Question → Ans*” fashion ([Chung et al., 2024](#)). This training objective encourages LLMs to reason based on either solely the context (CK) or their internal knowledge (PK) separately, rather than integrating both.

The method of new knowledge fabrication impacts recall of memory. Here, we further investigate how different methods (including *adding*, *variation*, and *dropping*) affect LLMs awareness and leverage of PK, by experimenting representative LLMs on ALCUNA in Conflicting Reasoning. This provides another view of the relation between

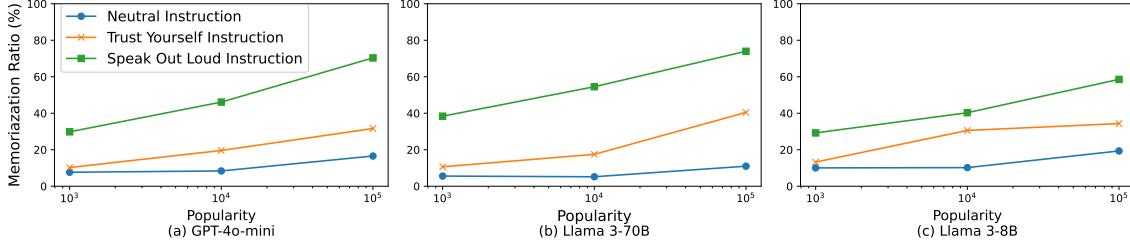


Figure 5: Memorization Ratio on ConflictQA across *popularity* categories by representative models, showing that LLMs recall their popular knowledge better.

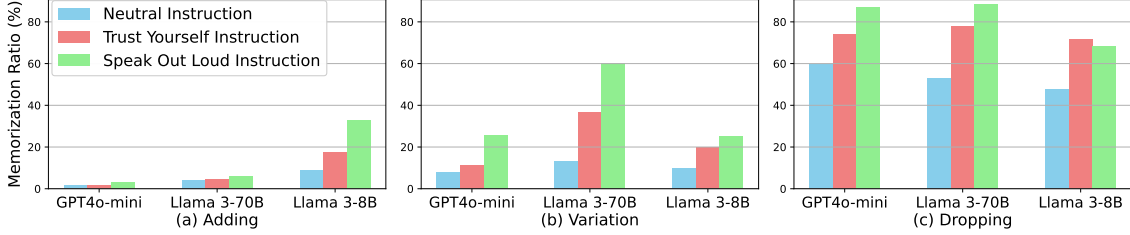


Figure 6: Memorization Ratio for Conflicting Reasoning under different ways of fabricating CK on ALCUNA. LLMs recall their knowledge way better with dropping than variation and adding.

CK and PK. Figure 6 provides the following insights:

First, *a significant gap exists between different fabrication methods*, although the overall trend observed in previous experiments remains consistent. This shows that LLMs awareness of PK is strongly related to how CK is fabricated from the PK.

Second, *LLMs are highly receptive to the new information (i.e., properties added to an existing entity)*, exhibiting mostly lower than 10% Memorization Ratio. This raises safety concerns, as LLMs may be convinced of the addition of disinformation to known entities, *e.g.*, fake news.

Third, *dropping of known facts does less harms in LLMs memory recall*. In Figure 6(c), with only “Neutral Instruction”, LLMs can leverage PK in up to 60% of cases, echoing results in Irrelevant Reasoning (Section 4.3) where LLMs recall their memories better than other reasoning types. However, they are still far from fully echoing their PK. We further discuss models ability to filter relevant information in Appendix B.5.

5.2 How to Better Leverage Parametric Knowledge Given the Context

As demonstrated above, LLMs struggle to integrate their PK with CK, which is crucial for both artificial and human intelligence. We have designed instructions to elicit PK to some extent (Section 3.2). Although effective, such progressively-enforced instructions are not for practical applications. The key challenge is to enable LLMs to naturally learn how to balance the two knowledge sources. A promising direction is to develop an agent frame-

work where LLMs assume a controlled amount of responsibility. By separating the process into knowledge recall and reasoning, LLMs may not be distracted by the context. We show preliminary explorations in Appendix B.6. Another potential solution is to fundamentally integrate CK and PK in post-training, which we leave for future work.

6 Conclusion

In this work, we comprehensively investigate how well LLMs can leverage their parametric knowledge (PK) in the presence of contextual knowledge (CK). Our contributions can be summarized:

- We formulate four *reasoning types*: *Supportive*, *Complementary*, *Conflicting* and *Irrelevant*, based on diverse relationships between CK and PK. To facilitate the study, we introduce a new dataset ECHOQA spanning scientific, factual and common-sense knowledge for multiple LLMs.

- We find that LLMs consistently struggle to fully leverage their PK when CK is present under different reasoning types. We design a series of *progressively-enforced instructions* to guide LLMs leverage of PK. While the instructions improve their awareness of PK, they still fall short of expectations.

- We further show that LLMs awareness of their PK can be affected by the relations between CK and PK, instruction, entity popularity and methods of fabricating CK, showing the vulnerability when faced with external knowledge.

We hope that our work provides valuable insights and resources to guide future explorations in understanding LLM-based systems.

Limitations

Our study reveals a key vulnerability in current LLMs: they struggle to fully leverage their parametric knowledge when contextual knowledge is present. However, there are some limitations to our work. First, we base our study primarily on contemporary LLMs with strong reasoning and instruction-following abilities. While we find evidence that this issue can be partially mitigated through dedicated prompt or framework design, we argue that the ability to effectively integrate PK with CK should be inherent to the base model. We aim for improvements in this area in future work. Second, while we consistently observe our findings across various reasoning types, we did not focus on making the data more challenging for broader coverage. For instance, conflicting knowledge could occur at the entity, sentence, or document level. A deeper exploring on such distinctions is left for future research. Third, real-world applications are likely more complex than the reasoning types we have designed. Although we categorize the interactions between PK and CK into different reasoning types, there may be more nuanced combinations of these relationships in real-world scenarios. We hope our findings raise awareness of the safety concerns surrounding the trustworthiness and reliability of LLM-based systems.

Acknowledgments

References

- Zeyuan Allen-Zhu and Yuanzhi Li. 2023. Physics of language models: Part 3.1, knowledge storage and extraction. *arXiv preprint arXiv:2309.14316*.
- Kaikai An, Fangkai Yang, Liquan Li, Junting Lu, Sitao Cheng, Lu Wang, Pu Zhao, Lele Cao, Qingwei Lin, Saravan Rajmohan, Dongmei Zhang, and Qi Zhang. 2024. [Thread: A logic-based data organization paradigm for how-to question answering with retrieval augmented generation](#). *Preprint*, arXiv:2406.13372.
- Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. [Self-RAG: Learning to retrieve, generate, and critique through self-reflection](#). *arXiv preprint arXiv:2310.11511*.
- Sitao Cheng, Ziyuan Zhuang, Yong Xu, Fangkai Yang, Chaoyun Zhang, Xiaoting Qin, Xiang Huang, Ling Chen, Qingwei Lin, Dongmei Zhang, Saravan Rajmohan, and Qi Zhang. 2024. [Call me when necessary: LLMs can efficiently and faithfully reason over structured environments](#). In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4275–4295, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.
- Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. [Measuring and improving consistency in pretrained language models](#). *Transactions of the Association for Computational Linguistics*, 9:1012–1031.
- Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. [Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies](#). *Transactions of the Association for Computational Linguistics*, 9:346–361.
- Eleanor J Gibson. 1988. Exploratory behavior in the development of perceiving, acting, and the acquiring of knowledge. *Annual review of psychology*, 39(1):1–42.
- Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML’20*. JMLR.org.
- Xiang Huang, Sitao Cheng, Yuheng Bao, Shanshan Huang, and Yuzhong Qu. 2023. [MarkQA: A large scale KBQA dataset with numerical reasoning](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10241–10259, Singapore. Association for Computational Linguistics.
- Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. [Active retrieval augmented generation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.
- Mingyu Jin, Weidi Luo, Sitao Cheng, Xinyi Wang, Wenye Hua, Ruixiang Tang, William Yang Wang, and Yongfeng Zhang. 2024. [Disentangling memory and reasoning ability in large language models](#). *Preprint*, arXiv:2411.13504.
- Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d’Autume, Tomáš Kočiský, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. 2021. [Mind the gap: Assessing temporal generalization in neural language models](#). In *Advances in Neural Information Processing Systems*.

- Zhaoyi Li, Gangwei Jiang, Hong Xie, Linqi Song, Defu Lian, and Ying Wei. 2024. [Understanding and patching compositional reasoning in LLMs](#). In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9668–9688, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.
- Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. [Entity-based knowledge conflicts in question answering](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. [When not to trust language models: Investigating effectiveness of parametric and non-parametric memories](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.
- Meta. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. [Can a suit of armor conduct electricity? a new dataset for open book question answering](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.
- Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.
- OpenAI. 2023. Gpt-4o system card. <https://openai.com/index/gpt-4o-system-card/>. Accessed: 2024-09-30.
- OpenAI. 2024. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- OpenAI. 2024. Introducing openai o1-preview. <https://openai.com/index/introducing-openai-o1-preview/>. Accessed: 2024-09-30.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. 2023. [Measuring and narrowing the compositionality gap in language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711, Singapore. Association for Computational Linguistics.
- Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Yufei Huang, Chaojun Xiao, Chi Han, Yi Ren Fung, Yusheng Su, Huadong Wang, Cheng Qian, Runchu Tian, Kunlun Zhu, Shi Liang, Xingyu Shen, Bokai Xu, Zhen Zhang, Yining Ye, Bo Li, Ziwei Tang, Jing Yi, Yu Zhu, Zhenning Dai, Lan Yan, Xin Cong, Ya-Ting Lu, Weilin Zhao, Yuxiang Huang, Jun-Han Yan, Xu Han, Xian Sun, Dahai Li, Jason Phang, Cheng Yang, Tongshuang Wu, Heng Ji, Zhiyuan Liu, and Maosong Sun. 2023. [Tool learning with foundation models](#). *ArXiv*, abs/2304.08354.
- Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. [Toolformer: Language models can teach themselves to use tools](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Alon Talmor and Jonathan Berant. 2018. [The web as a knowledge-base for answering complex questions](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 641–651, New Orleans, Louisiana. Association for Computational Linguistics.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *CoRR*, abs/2307.09288.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multi-hop questions via single-hop question composition.

- Transactions of the Association for Computational Linguistics*, 10:539–554.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. [Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037, Toronto, Canada. Association for Computational Linguistics.
- Boshi Wang, Xiang Yue, Yu Su, and Huan Sun. 2024. [Grokking transformers are implicit reasoners: A mechanistic journey to the edge of generalization](#). *Preprint*, arXiv:2405.15071.
- Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xian-gru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, et al. 2023a. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *arXiv preprint arXiv:2310.07521*.
- Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2023b. Resolving knowledge conflicts in large language models. *arXiv preprint arXiv:2310.00935*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. [Chain of thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*.
- Siye Wu, Jian Xie, Jiangjie Chen, Tinghui Zhu, Kai Zhang, and Yanghua Xiao. 2024. How easily do irrelevant inputs skew the responses of large language models? *arXiv preprint arXiv:2404.03302*.
- Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. [Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts](#). In *The Twelfth International Conference on Learning Representations*.
- Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. [Knowledge conflicts for llms: A survey](#). *Preprint*, arXiv:2403.08319.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.
- Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. 2024b. [Do large language models latently perform multi-hop reasoning?](#) In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10210–10229, Bangkok, Thailand. Association for Computational Linguistics.
- Xunjian Yin, Baizhou Huang, and Xiaojun Wan. 2023. [ALCUNA: Large language models meet new knowledge](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1397–1414, Singapore. Association for Computational Linguistics.
- Hao Zhang, Yuyang Zhang, Xiaoguang Li, Wenxuan Shi, Haonan Xu, Huanshuo Liu, Yasheng Wang, Lifeng Shang, Qun Liu, Yong Liu, et al. 2024. Evaluating the external and parametric knowledge fusion of large language models. *arXiv preprint arXiv:2405.19010*.
- Zexuan Zhong, Zhengxuan Wu, Christopher Manning, Christopher Potts, and Danqi Chen. 2023. [MQuAKE: Assessing knowledge editing in language models via multi-hop questions](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15686–15702, Singapore. Association for Computational Linguistics.
- Ziyuan Zhuang, Zhiyang Zhang, Sitao Cheng, Fangkai Yang, Jia Liu, Shujian Huang, Qingwei Lin, Saravan Rajmohan, Dongmei Zhang, and Qi Zhang. 2024. [Efficientrag: Efficient retriever for multi-hop question answering](#). *Preprint*, arXiv:2408.04259.

Appendix

In the following sections, we detail on the following aspects:

- Appendix A: Details of Data Constructions
- Appendix B: Details of Experiments
- Appendix C: Prompt Examples

A Elaboration on Data Construction

As illustrated in Section 3.3, our proposed dataset **ECHOQA** aims to test LLMs awareness and leverage of parametric knowledge (PK) given contextual knowledge (CK). EchoQA is adapted from existing datasets, the construction procedure of which can fit into any knowledge-intensive task. The key step is to obtain LLMs PK then construct CK and answer choices to fit our reasoning types, as in Figure 1. By default, the answer choice includes the answer sticking to CK and PK for Conflicting Reasoning or the ground truth for other reasoning types, a randomly-sampled option with the same type, and an unknown option. We conduct human evaluation for the generated questions and contextual knowledge to ensure the meaningfulness, knowledge faithfulness and fluency. Examples for each dataset is shown in Table 12, 13, 14, *w.r.t.* our reasoning types. Statistics for each LLM are shown in Table 4 Here we show more details.

ALCUNA (Yin et al., 2023) originally provides new knowledge. To obtain PK, we adopt a multi-choice question about the fact required by the original question. For example, “Does cat prey on fish? A. Yes, B. No, C. Unknown” or “Cat prey on what? A. Fish, B. Tiger, C. Unknown”. The exact fact for the question is provided by the dataset. The knowledge association subset naturally fits our Complementary Reasoning Type. Questions for other reasoning types are sampled from the knowledge differentiation subset. For Conflicting Reasoning, we regard the new knowledge adapted from the old one as a conflict fact by changing the new entity name back. For example, given “Dogcat prey on rabbit”, we change it back to “Cat prey on rabbit” as a conflict fact. The entity name in the question also need to change accordingly. For Irrelevant Reasoning, we randomly sample 1-4 new entities profiles provided as CK.

ConflictQA (Xie et al., 2024) originally provides conflicting factual knowledge based on PopQA (Mallen et al., 2023) and StrategyQA (Geva

et al., 2021). We adopt the PopQA subset, with the *popularity* of a question, counted by monthly Wikipedia page views associated with entities mentioned in the question. The key challenge is to construct new knowledge never seen by the LLMs for Complementary Reasoning. Following Mallen et al., 2023, we adopt templates for each relationship (Table 11). With the template, we obtain CK and construct new multi-hop questions. For example, for question “What’s the occupation of Michael Jordan”, we adopt template to create a new entity “PersonA” with profile “Michael Jordan is the famous friend of PersonA”. And we construct the question “What’s the occupation of the famous friend of PersonA”. For Irrelevant Reasoning, we randomly sample provided knowledge with the same relationship.

MuSiQue (Trivedi et al., 2022) provides all supporting paragraphs, from which we can query LLMs to obtain the PK and CK, in a manner similar to ALCUNA. The sub-questions and documents can be directly used for Complementary Reasoning. We also adopt the answers of sub-questions as answer choices for Complementary Reasoning. The key challenge for this dataset is to create conflicting knowledge. We prompt Llama 3.1-70B by zero-shot to generate conflicting knowledge and a related question, based on the filtered PK, for Conflicting Reasoning. The prompt is shown in Table 17. The constructed questions are also used for Irrelevant Reasoning, where we sample irrelevant documents from the provided unused paragraphs to test factual knowledge with a longer context.

OpenBookQA (Mihaylov et al., 2018) provides some scientific facts. To obtain CK and PK, we query LLMs to answer the question in a Chain-of-Thought (Wei et al., 2022) manner (Refer to Table 18 for prompt examples). We thereby obtain the PK by filtering the thoughts from items with the correct answers. Then, we prompt Llama 3.1-70B with one demonstration to generate complementary knowledge and questions (Refer to Table 18 for prompt examples). The Conflicting and Irrelevant Reasoning data construction is similar a MuSiQue.

B Elaboration on Experiments

B.1 Language Model API/Checkpoints

This section elaborates on checkpoints we adopt for experiments. All open-source models are available on the Hugging Face platform. For Llama 3.1, we use “meta-llama/Llama-3.1-(8,

Models	ALCUNA (#)	ConflictQA (#)	MuSiQue (#)	OpenBookQA (#)
OpenAI o1	537/512/1,119	1,993	611	488
GPT-4o	631/740/1851	3,001	525	476
GPT-4o-mini	599/852/1,811	1,801	675	458
Llama 3.1-70B	1,090/550/1,809	2,372	635	397
Llama 3.1-8B	905/873/1,953	2,299	393	351
Qwen 2-7B	993/346/1,397	2,143	452	415

Table 4: Number of examples for each LLM for Complementary/Conflicting/Irrelevant reasoning type. The difference between LLMs is due to their different possession of knowledge.

Model	A	A*	C	C*	M	M*
GPT-4o	51.94 (33.74)	91.97 (8.08)	91.13 (6.14)	98.20 (1.07)	76.76 (8.62)	82.10 (7.43)
GPT-4o-mini	36.33 (58.57)	87.15 (12.52)	79.50 (12.45)	97.75 (0.87)	75.84 (11.07)	79.41 (9.33)
Llama 3.1-70B	50.27 (11.38)	91.30 (0.08)	82.82 (5.92)	96.79 (1.87)	72.87 (9.68)	80.88 (7.72)
Llama 3.1-8B	33.37 (32.15)	82.87 (12.70)	65.59 (9.80)	92.04 (5.79)	33.86 (8.25)	48.79 (8.86)
Qwen 2-7B	28.97 (35.25)	68.18 (28.60)	31.17 (62.20)	74.94 (20.49)	30.75 (30.31)	45.13 (20.80)

Table 5: Accuracy (Unknown Ratio) of “Speak Out Loud” Instruction for Complementary Reasoning on ALCUNA. A, C, M denotes ALCUNA, ConflictQA and MuSiQue, respectively. * denotes Golden Knowledge is given in the context to show the gap between the performance.

Model	Single-hop	Multi-hop
GPT-4o-mini	98.97	87.15
Llama 3.1-70B	97.80	91.30
Llama 3.1-8B	93.29	82.87

Table 6: Accuracy for Supportive Reasoning, with single- and multi-hop questions on ALCUNA.

70B-Instruct”, respectively. For Qwen 2, we use “Qwen/Qwen2-7B-Instruct”. We use OpenAI o1-preview, GPT-4o and GPT-4o-mini from OpenAI platform (<https://platform.openai.com>).

B.2 Results for Supportive Reasoning

Here we report the results for Supportive Reasoning on ALCUNA with representative LLMs. We do not include it in main experiments for two reasons: 1) Intuitively, models perform quite well in this type. As we make share the PK is held by LLMs and CK represents the same knowledge, the result is straightforward; 2) More importantly, it is hard to tell whether LLMs use CK or PK for reasoning when CK is present. We test questions require single and multiple facts for reasoning, as shown in Table 6. It is demonstrated that LLMs performance quite well with this type, reaching our expectation. However, based on our previous conclusions that LLMs

might excessively rely on the context for reasoning, so they still may not perfectly leverage their own knowledge.

B.3 Results of “Speak Out Loud” Instruction for Complementary Reasoning on ALCUNA

Due to space limits, we do not cover the results of “Speak Out Loud” Instruction for Complementary Reasoning on ALCUNA in the main part of our paper. Here we show the detailed results which also sticks to our conclusion that even though the CK is complementary and explicitly asked to use their PK, LLMs still suppress their PK to a great extend.

B.4 Analysis of Models Uncertainty

Here we report results by Unknown Ratio U_R , which holds similar trend across all datasets and all tested LLMs. Therefore, we do not include all of them in the main body of our paper. Figure 7 illustrates the U_R for Complementary Reasoning. Similarly, Table 7, 8 and Figure 8 shows the U_R for Conflicting and Irrelevant Reasoning, respectively.

Echoing discussion in the main body (Section 4.1), the introduction of complementary contextual knowledge may even bring up the uncertainty of models (Figure 7 (a)). With our progressively-

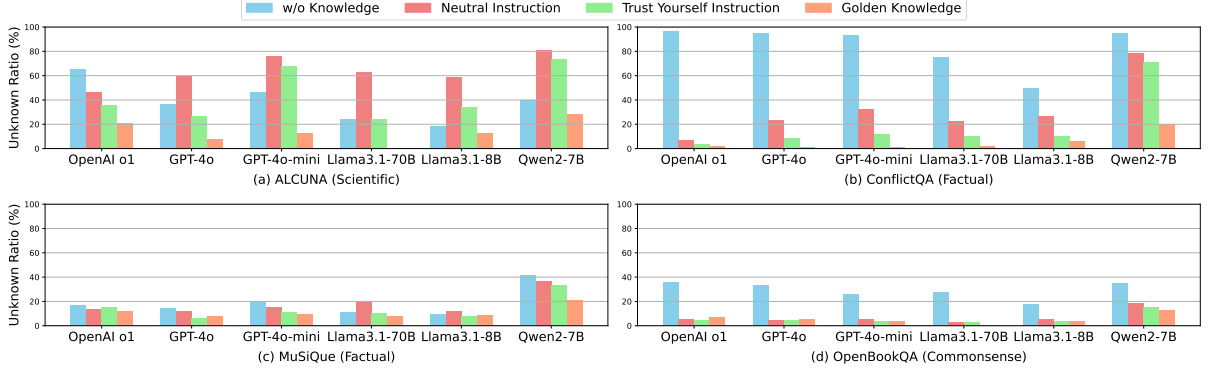


Figure 7: Unknown Ratio for Complementary Reasoning across four datasets. w/o Knowledge and Golden Knowledge means no information and all the required information is given, respectively. The x -axis denotes different models.

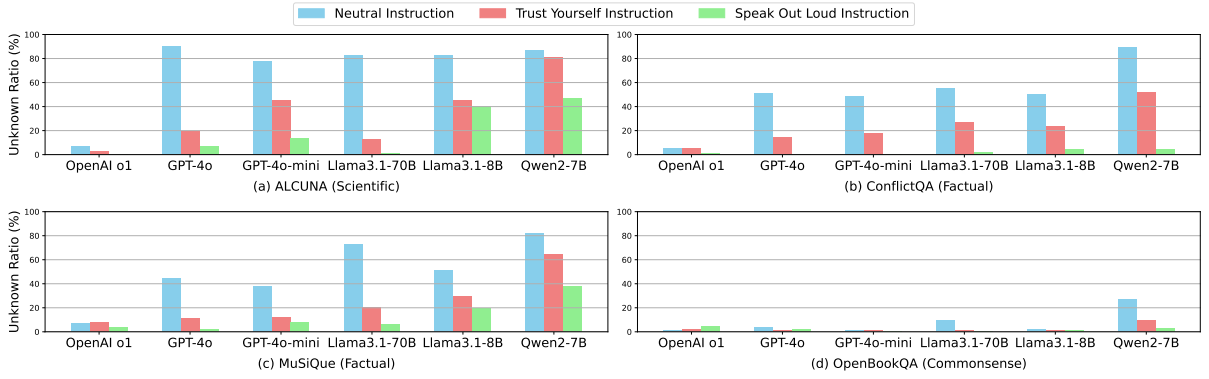


Figure 8: Unknown Ratio for Irrelevant Reasoning.

enforced Reasoning Instructions, LLMs uncertainty go down substantially. But still, although given the sufficiency of knowledge, LLMs tend to choose “unknown” to some extent. This again echoes our finding that LLMs cannot fully leverage their knowledge when CK is present. Also, the U_R for commonsense knowledge is quite low, compared with other types of knowledge, shedding lights on the future improvements.

For Conflicting Reasoning as shown in Table 7, 8, stronger models show relative lower uncertainty (lower than 7%). For weaker models in scientific knowledge, such as Llama 3.1-8B and Qwen 2-7B, while our progressively enforced instruction improves their leverage of PK and encouraging them to trust themselves, the U_R also goes up slightly. We speculate that while the absolute rate of trusting themselves increases, LLMs cannot fully recall their memory of the scientific knowledge distracted by the presence of CK, even when asked to output their own knowledge.

For Irrelevant Reasoning as shown in Figure

8, given the Neutral Instruction, LLMs tend to lean on context for reasoning, showing high uncertainty. When the context is irrelevant, although they have the knowledge to answer the question, they still tend to respond “Unknown”, with Qwen 2-7B reaching 80% U_R . Concerningly, our progressively enforced instructions lower the uncertainty by a large margin, with “Speak Out Loud Instruction” almost dropping U_R to lower than 10% for most LLMs. Also, LLMs exhibit less uncertainty when faced with commonsense knowledge, echoing previous conclusions that LLMs may recall this type of knowledge better than others.

B.5 When CK is Required and Mixed with Irrelevant Information

As we focus on investigating LLMs ability to leverage PK, we do not include when PK is irrelevant to the question while the required CK is mixed with irrelevant knowledge. However, it is also an inevitable case for RAG-systems and shows LLMs ability of *knowledge discrimination or filtering*. We

Model	NI	TYI	SOLI
OpenAI o1	4.10	16.01	10.93
GPT-4o	6.22	3.10	6.48
GPT-4o-mini	2.34	1.76	1.76
Llama 3.1-70B	5.18	2.36	6.00
Llama 3.1-8B	10.99	13.51	30.58
Qwen 2-7B	21.38	16.47	19.36

Table 7: Unknown ratio U_R for Conflicting Reasoning on ALCUNA. NI, TYI, SOLI denotes Neutral, Trust Yourself, Speak Out Loud Instructions, respectively.

Model	NI	TYI	SOLI
OpenAI o1	7.29	5.13	3.49
GPT-4o	3.33	3.93	4.39
GPT-4o-mini	3.06	4.22	0.78
Llama 3.1-70B	1.98	1.99	3.02
Llama 3.1-8B	3.09	5.87	9.40
Qwen 2-7B	4.56	3.36	8.35

Table 8: Unknown ratio U_R for Conflicting Reasoning on ConflictQA. NI, TYI, SOLI denotes Neutral, Trust Yourself, Speak Out Loud Instructions, respectively.

Model	Strategy	Complementary	Conflicting	Irrelevant
Llama 3.1-70B	Neutral Instruction	27.51	13.15	15.87
	Agent Framework	34.62	58.09	86.11
Llama 3.1-8B	Neutral Instruction	23.61	9.56	10.14
	Agent Framework	39.41	32.44	45.77
Qwen 2-7B	Neutral Instruction	11.08	15.13	2.29
	Agent Framework	23.23	53.22	61.46

Table 9: Results of Neutral Instruction and Agent Framework with various reasoning types on ALCUNA.

Table 10: Accuracy for Irrelevant Reasoning, requiring CK on ALCUNA. Single and w/ Irr Fact denotes context with only the required fact and with randomly-sampled 1-4 irrelevant facts, respectively.

Model	Single Fact	w/ Irr Fact
GPT-4o-mini	79.78	78.81
Llama 3.1-70B	86.81	87.75
Llama 3.1-8B	71.13	65.56

experiment on ALCUNA with the knowledge understanding subset, making sure the knowledge required for the question is not held by LLMs. Results are shown in Table 10. It is demonstrated that for knowledge new to the LLMs, their understanding ability is only decent, with Llama 3.1-8B reaching 71.13% Accuracy. However, with irrelevant documents, the result goes even higher for some models, showing that the reasoning ability is not greatly affected by the irrelevant knowledge. Compared to the main results in Figure 4, such result indicates that *LLMs can filter relevant information effectively in the context, but cannot disentangle their own knowledge if it is required for the question.*

B.6 A Possible Solution to Better Leverage Parametric Knowledge: Agent

The main focus of our paper is to investigate how well LLMs can leverage their parametric knowledge (PK) when various kinds of contextual knowledge (CK) is present. We observe that LLMs tend to suppress their PK regardless of models, instructions and knowledge types. In the main part of our paper, we only discuss some possible solutions to fix this problem. Here we conduct preliminary experiments with an agent framework combined of a knowledge agent and a reasoning agent.

Given a question Q , the knowledge agent generates knowledge PK relevant to Q . Then, the reasoning agent reasons based on Q , PK and the given contextual knowledge CK to generate the final answer. With two agents assuming controlled amount of responsibility, the knowledge agent conditioned solely on the question may not be distracted by the context. And the reasoning agent can reason based directly on the information in the context. Note that such agent framework can be used in real-world practical applications, which is more natural compared to our designed progressively-enforced instructions (Section 3.2). Moreover, such framework can be further used to fine-tune existing LLMs to enhance their ability of leveraging their on knowledge (Jin et al., 2024).

We experiment with representative LLMs on ALCUNA with our designed reasoning types. Still, we adopt zero-shot prompting for all tested LLMs. Other settings remain the same as the main experiments. As illustrated in Figure 9, the agent framework significantly enhances LLMs utilization of their parametric knowledge for all reasoning types. Specifically, the agent framework almost doubles the accuracy of complementary reasoning for Llama 3.1-8B and Qwen 2-7B. It also improving the recall of parametric knowledge from 10.14% and 2.29% to 45.77% and 61.46% for Llama 3.1-8B and Qwen 2-7B, respectively. However, although with decent performance of the agent framework, there still exists much room for improvement. Another promising way is to introduce robust post-training methods, which we leave for future work.

B.7 Case Study

To figure out why LLMs suppress their own knowledge at the presence of contextual knowledge. We conduct case study over 200 random samples generated by Llama 3.1-70B with Neutral Instruction (Section 3.2) on ALCUNA (Yin et al., 2023) where LLMs were able to answer on their own but failed when complementary knowledge was introduced. Note that we do ask LLMs to use their own knowledge in the instruction (Section 3.2). We discover that in 93.5% of cases, LLMs solely rely on the context and even identified the inefficiency of knowledge for reasoning, as an example shown in Table 15. Without contextual knowledge, the LLMs can infer the new entity by its name, *e.g.*, *dogcat might have properties similar to dog and cat*. Then, LLMs can reason by the process of elimination to figure out the correct answer. However, the introduce of CK is likely to limit LLMs reasoning ability to some extent in scientific knowledge. In other cases, the LLMs try to recall their own knowledge, but is eventually distracted by the options or choose the wrong option by hallucination, as examples shown in Table 16.

C Prompt Design

We provide a comprehensive list of all the prompts that have been utilized in this study, offering a clear reference for the understanding our experimental approach.

Specifically, in Table 17, 18 we provide prompt examples adopted to construct ECHOQA. In Table

C, we show some prompt templates of our designed progressively-enforced instructions for various reasoning types (Section 3.2). Note that for all instructions, we ask GPT-4o for paraphrasing to eliminate the possible bias of instructions. In Table 19, we provide prompt examples for our designed Reasoning Instructions.

Examples of Progressively-Enforced Reasoning Instructions

Neutral: Combine the given information and your OWN knowledge to answer questions.

Trust Yourself: The given information is NOT SUFFICIENT, you should USE your OWN knowledge combined with the information to answer questions. (Complementary)

If you think the given information is CONFLICTED with your knowledge, you should USE your OWN knowledge to answer questions. (Conflicting)

Speak Out Loud: You must first OUTPUT your OWN knowledge about the question. If you think the given information is CONFLICTED with your knowledge, should USE your OWN knowledge to answer questions. (Conflicting)

Relationship	Template
Occupation	the person who frequently collaborates with [subj] the best friend of [subj] the famous friend of [subj]
Place of Birth	the renowned figure who is a friend of [subj] the best friend of [subj] the person who frequently collaborates with [subj]
Genre	the influential entity previously related with [subj]
Father	the renowned figure with [subj] the famous person known by [subj] the best friend of [subj]
Country	the renowned entity similar to [subj] the notable figure similar to [subj]
Producer	the renowned product similar to [subj] the famous product similar to [subj] the prominent product like [subj]
Director	the renowned film similar to [subj] the famous film similar to [subj] the renowned movie similar to [subj]
Capital of	the famous place near [subj] the renowned place near [subj] the prominent place similar to [subj]
Screenwriter	the renowned screenplay similar to [subj] the famous TV shows similar to [subj] the prominent show like [subj]
Composer	the work similar to [subj] the prominent work similar to [subj] the influential work similar to [subj]
Color	the notable entity similar to [subj] the prominent entity similar to [subj]
Religion	the best friend of [subj] the prominent friend of [subj]
Sport	the prominent entity related to [subj] the famous entity similar to [subj]
Author	the prominent work similar to [subj] the influential work similar to [subj]
Mother	the renowned figure with [subj] the famous person known by [subj] the best friend of [subj]
Capital	the important place near [subj] the big place near [subj] the neighbor place of [subj]

Table 11: Templates to connect new knowledge to parametric knowledge for PopQA. [subj] denotes subject entity in the new question. For example, for question “*What’s the occupation of Michael Jordan*”, we adopt template to create a new entity with profile “*Michael Jordan is the famous friend of PersonA*”. And we change the question to “*What’s the occupation of the famous friend of PersonA*”.

ALCUNA

Question: What type of flowers are visited by an organism that is preyed upon by the co-roosting partners of 'Myotis lucifralis'?

0. Haustrium 1. Picea sitchensis 2. Scalariogyra 3. Stellaria media

Contextual Knowledge: Myotis lucifralis co-roost with Myotis nattereri

Ground Truth: 1

ConflictQA

Question: What is the person who frequently collaborates with PersonA's occupation?

0. cartoonist, illustrator 1. record producer 2. film director 3. unknown

Contextual Knowledge: Eleanor Davis is the person who frequently collaborates with PersonA.

Ground Truth: 0

MuSiQue

Question: What administrative territorial entity is the owner of Ciudad Deportiva located?

0. Tamaulipas 1. Nuevo Laredo 2. unknown

Contextual Knowledge: The Ciudad Deportiva ("Sports City") is a sports complex in Nuevo Laredo, Mexico....

Ground Truth: 0

OpenbookQA

Question: Which activity can help a person save money on lunch expenses if they usually eat lunch out?

Contextual Knowledge: DailyBite offers lunch coupons to frequent customers. FoodieClub is an expensive membership-based dining group...

Ground Truth: 0

Table 12: Complementary Reasoning Examples in ECHOQA

ALCUNA

Question: What's the longitude of the location where Phcytodidae is commonly found?

0. 1.61505 degrees 1. 93.05 degrees 2. unknown

Contextual Knowledge: Phcytodidae longitude 1.61505 degrees...

A_{ck} : 0, A_{pk} : 1

ConflictQA

Question: Who was the screenwriter for The Hunt?

0. Nick Cuse and Damon Lindelof 1. David Judah Simon 2. unknown

Contextual Knowledge: David Judah Simon is a highly acclaimed screenwriter

A_{ck} : 1, A_{pk} : 0

MuSiQue

Question: Was Green recorded as a standalone album after the US tour in late 1977? 0. Yes 1. No 2. unknown

Contextual Knowledge: Green was not recorded alone, but rather in conjunction with Motivation Radio..

A_{ck} : 1, A_{pk} : 0

OpenbookQA

Question: In what type of environments is fog more likely to form?

0. Environments with higher humidity, such as marshes. 1. Environments with lower humidity, such as deserts. 2. unknown

Contextual Knowledge: Fog is more likely to form in dry environments with low humidity levels.

A_{ck} : 1 A_{pk} : 0

Table 13: Conflicting Reasoning Examples in ECHOQA. A_{ck} and A_{pk} denotes answer sticking to contextual and parametric knowledge, respectively.

ALCUNA

Question: Does the longitude of Phcytodidae equal 1.61505 degrees?

0. yes 1. no 2. unknown

Contextual Knowledge: Argasidae longitude 88.54616346104861 degrees... Lichida latitude 71.4882053945494 degrees ...

Ground Truth: 1

ConflictQA

Question: Who is the author of The Program?

0. Suzanne Young. 1. Gregg Hurwitz 2. unknown

Contextual Knowledge: Anna Todd is a bestselling author who gained popularity...Cormac McCarthy is a highly acclaimed American novelist and playwright...

Ground Truth: 0

MuSiQue

Question: Where was the Philae lander operated from?

0. DLR's Lander Control Center in Cologne, Germany 1. A secret facility in the United States 2. unknown

Contextual Knowledge: Empire Sports Network was an American regional sports network...Nova Sports is a premium Sports service in Greece and is owned by Forthnet Group...

0. Eating lunch at DailyBite and using their coupons 1. Having lunch with FoodieClub 2. unknown

Ground Truth: 0

OpenbookQA

Question: What is the primary source of sustenance for predators in an ecosystem? 0. Herbivores or other animals~ 1. Plants and fung 2. unknown

Contextual Knowledge: Tree roots can push through obstacles like sidewalks as they grow...

Ground Truth: 0

Table 14: Irrelevant Reasoning Examples in ECHOQA

Question: What is the prey of the organism that shares a roosting place with the species 'Myotis lucifralis'?

0. Chara andina 1. Aldabrachelys 2. Noctuidae 3. Geomyidae

Ground Truth: 2

Output w/o Contextual Knowledge:

The species 'Myotis lucifugus' (little brown bat) shares a roosting place with 'Myotis lucifugus', so the correct species to consider is 'Myotis lucifugus'. Now, we need to determine the prey of 'Myotis lucifugus'. Bats are insectivores, and their diet consists of various insects. Among the given options, 'Noctuidae' is a family of moths, which are known to be preyed upon by bats.

Final Choice: 2

Output w/ Contextual Knowledge: The organism that shares a roosting place with the species 'Myotis lucifralis' is 'Myotis nattereri'. Since the provided information about 'Myotis lucifralis' does not include the prey of 'Myotis nattereri', but we know 'Myotis lucifralis' preys on 'Noctuidae', among others, and it is likely that 'Myotis nattereri' has a similar diet, but without the exact information about 'Myotis nattereri', the best answer is 'Unknown' since we can't be certain what 'Myotis nattereri' preys on. However, we can make an educated guess that it would be similar to 'Myotis lucifralis', but the options given do not include any prey of 'Myotis lucifralis' other than 'Noctuidae' which is a possible answer if we were certain they had the same prey. In this case, the best answer is 'Unknown' or 'Noctuidae' if we are guessing.

Final Choice: 4

Question: What is the species that Myotis lucifralis co-roosts with through three intermediate species?

0. Annona pruinosa 1. Plecotus auritus 2. Tuxedo drakei 3. Struvea thoracica

Ground Truth: 1

Output w/o Contextual Knowledge:

Myotis lucifugus, also known as the little brown bat, is a species of bat known to co-roost with other species. However, the question is specifically asking about co-roosting through three intermediate species. Given the options, I must consider which species is biologically plausible to co-roost with Myotis lucifugus through intermediate species. Plecotus auritus, also known as the brown long-eared bat, is known to co-roost with other bat species. Based on the information available, I consider it the most likely candidate among the options provided.

Final Choice: 1.

Output w/ Contextual Knowledge:

The given information about Myotis lucifralis only mentions that it co-roosts with Myotis nattereri. However, the question asks for a species that Myotis lucifralis co-roosts with through three intermediate species. Since there is no information provided about the co-roosting relationships of Myotis nattereri or any other potential intermediate species, we cannot determine the correct answer based on the given information. Therefore, the correct choice is "Unknown".

Final Choice: 4

Table 15: Examples where complementary knowledge drops the performance on ALCUNA. The LLMs reply excessively on context in these examples.

Question: What is eaten by the organism that is parasitized by 'Argasidae', and is itself a predator of another species? 0. Dasypus novemcinctus 1. Ochropleura candelisequa 2. Danaus plexippus 3. Anthurium fragrans 4. Unknown

Contextual Knowledge: Argasidae parasitize Homo sapiens.....

Ground Truth: 0

Output w/o Contextual Knowledge: Argasidae is a family of ticks, commonly known as soft ticks. These ticks are ectoparasites that feed on the blood of mammals and birds. Given the options, we need to find an organism that is a predator of another species and can be parasitized by Argasidae. Dasypus novemcinctus, also known as the nine-banded armadillo, is a predator that feeds on various invertebrates and small vertebrates. It can be parasitized by ticks, including Argasidae. Ochropleura candelisequa is a moth, Danaus plexippus is a butterfly, and Anthurium fragrans is a plant, none of which are likely to be parasitized by Argasidae in the context of being a predator of another species. Final Choice: 0

Output w/ Contextual knowledge: The question states that the organism is parasitized by 'Argasidae' and is itself a predator of another species. From the given information, we know that 'Argasidae' parasitizes 'Procyon lotor'. We also know that 'Procyon lotor' is eaten by 'Argasidae'. However, the question asks for what is eaten by the organism that is parasitized by 'Argasidae', and is itself a predator of another species, which in this case is 'Procyon lotor'. Since we don't have information about what 'Procyon lotor' eats in the provided data, we can't give a specific answer from the options. However, 'Procyon lotor' is known to eat various small animals including insects, among them, some species of moths such as 'Ochropleura candelisequa' cousins, which is close to option 1. Final Choice: 1

Question: Uredo has a host, which in turn serves as a host for another organism. What is the host of this second organism? 0. Anthaxia 1. Pelophylax ridibundus 2. Lolium perenne 3. Tradescantia 4. Unknown

Contextual Knowledge: Uredo have host Avena sativa.....

Ground Truth: 2

Output w/o Contextual Knowledge: To answer this question, we need to break it down step by step. Uredo is a genus of rust fungi, which are known to be parasitic on plants. Given that Uredo has a host, which in turn serves as a host for another organism, we can infer that the host of Uredo is likely a plant. Considering the options provided, we should look for the one that is most likely to be a host for Uredo and then a host for another organism.

Looking at the options: - Anthaxia is a genus of jewel beetles, which are not typically hosts for rust fungi. - Pelophylax ridibundus is a species of frog, which is not a host for Uredo. - Lolium perenne is a species of ryegrass, a type of plant. - Tradescantia is a genus of plants commonly known as spiderworts.

Given that rust fungi typically infect plants, we can narrow down our options to Lolium perenne and Tradescantia. However, we need to consider which of these plants is more likely to be a host for another organism that also serves as a host for Uredo.

Considering the common hosts for Uredo and the possibility of another organism being hosted by the same plant, we should consider Lolium perenne. It is known that some rust fungi infect ryegrass, and this plant can also host other types of organisms.

Final Choice: 2

Output w/ Contextual Knowledge: The hosts of Uredo are listed in the provided information, which includes various plant species. Among the options given, we should look for a plant species that is known to be a host for another organism. Tradescantia is a plant genus that is known to be a host for various organisms, including fungi, insects, and other invertebrates. On the other hand, the other options do not directly relate to known hosts of Uredo or other organisms.

Final Choice: 3

Table 16: Examples where complementary knowledge drops the performance on ALCUNA. The LLMs inhibits deep thinking with contextual knowledge compared to ones without knowledge in these examples.

Operation	Prompts
Conflict Knowledge and Question Generation	<p>You are tasked with generating conflicting knowledge based on the given knowledge and creating a related multiple-choice question. Follow these steps:</p> <ol style="list-style-type: none"> 1. Generate Conflict Knowledge: Create a conflicting statement or fact that contradicts the given knowledge. This conflict should challenge the established understanding. 2. Formulate a Multiple Choice Question: Create a question that asks for the truth regarding the conflicting knowledge. The question should be about the conflict knowledge. Provide two answer choices. <p>Include:</p> <ol style="list-style-type: none"> 0. One choice that aligns with the given knowledge. 1. One choice that supports the conflicting knowledge. <p>Your generated conflicting knowledge should contain information with similar volume as the given knowledge. The Label of choice should be 0. and 1. Do not output anything after the Desired Output. Here is the knowledge you should refer: [Taxon] Desired Output Format: Thought: [thought] Conflicting Knowledge: [new_knowledge] Question: [new_question_with_two_choice] Let's think step by step.</p>

Table 17: Prompt Examples to construct conflicting knowledge and questions for ECHOQA.

Operation	Prompts
Commonsense Knowledge Elicitation	<p>You are an expert with factual commonsense knowledge on a wide range of topics. You will be given a multiple-choice question requiring commonsense knowledge to answer. Your task is to:</p> <ol style="list-style-type: none"> 1. Output the commonsense knowledge needed to answer the question, making it as concise as possible. 2. Output the answer based on the knowledge. <p>For the question and fact provided, provide the needed commonsense knowledge and the answer from the choices. Desired output format: Knowledge: ["knowledge1", "knowledge2", ...] Final Answer Choice: [Choice] DO NOT output anything after the Desired Output.</p> <p>Question: Which of these would let the most heat travel through? Choices: A. a new pair of jeans. B. a steel spoon in a cafeteria. C. a cotton candy at a store. D. a calvin klein cotton hat.</p> <p>Knowledge: ["Metal is a thermal conductor.", "Steel is made of metal.", "Heat travels through a thermal conductor."]</p> <p>Final Answer Choice: B</p> <p>Question: [question_stem] Choices: [formatted_choices] Knowledge:</p>
Complement Knowledge and Question Generation	<p>You are an expert with factual commonsense knowledge on a wide range of topics.</p> <p>Given a multiple-choice question, answer, and related commonsense knowledge, you must create a new question that requires both the given knowledge and new entities/knowledge to answer.</p> <p>Step 1: Generate New Entities Create new entities representing concepts or objects related to each choice. These entities must be new and not directly mentioned in the original context.</p> <p>Step 2: Create New Knowledge Create one-hop facts connecting the fake entities to the original choice. Answering the question must require both new and original knowledge. The new knowledge should be a list of strings.</p> <p>Desired Output Thought: [thought] New Question: [new_question] New Choices: [new_choices] New Knowledge: [new_knowledge] New Answer Key: [new_answer_key] Let's think step by step.</p>

Table 18: Prompt examples for knowledge and question construction for OpenBookQA.

w/o Knowledge	<p>You are a powerful question-answering system.</p> <p>Your task is to use your own commonsense knowledge to answer a multiple-choice question. If you are unable to answer the question, respond with the “Unknown” choice key.</p> <p>Answer the following question with your choice (A, B, C, D, etc.): [Question]</p> <p>Desired format: Thought process: [Thought process], Final Choice: [Final Choice].</p> <p>Let’s think step by step.</p>
Neutral / Golden Knowledge Instruction	<p>You are a powerful question-answering system.</p> <p>Users will provide some information along with a question. Your task is to combine the information provided by the user with your knowledge to answer the question. If you are unable to answer the question, respond with the “Unknown” choice key.</p> <p>Here is the basic information you can refer:[Taxon]</p> <p>Answer the following question with your choice (A, B, C, etc.): [Question]</p> <p>Desired format: Thought process: [Thought process], Final Choice: [Final Choice].</p> <p>Let’s think step by step.</p>
Trust Yourself Instruction	<p>You are a powerful question-answering system.</p> <p>Users will provide some information along with a question. Your task is to combine the information provided by the user with your knowledge to answer the question.</p> <p>If you are unable to answer the question, respond with the “Unknown” choice key.</p> <p>Here is the basic information you can refer: [Taxon]</p> <p>Answer the following question with your choice (A, B, C, etc.): [Question]</p> <p>Note that the information above is not sufficient to answer the question, you should use your own commonsense knowledge to further reason the final answer.</p> <p>Desired format: Thought process: [Thought process], Final Choice: [Final Choice].</p> <p>Let’s think step by step.</p>
Speak Out Loud Instruction	<p>You are a powerful question-answering system.</p> <p>Users will provide some information along with a question. Your task is to use the information provided by the user and your own knowledge to answer the question. If you are unable to answer the question, respond with the “unknown” choice.</p> <p>Answer the following question with your choice (A, B, C, etc.): [Question]</p> <p>You MUST first output your own knowledge about the question. I am sure you have the knowledge!</p> <p>Here is the Given Knowledge: [Taxon]</p> <p>Determine whether the given information about the question conflicts with your own knowledge.</p> <p>If you think the Given Knowledge conflicts with your own knowledge, you MUST USE your own knowledge to answer the question.</p> <p>Desired format: Thought process: [Thought process], My Own Knowledge: [My Own Knowledge], Final Choice: [Final Choice].</p> <p>Let’s think step by step.</p>

Table 19: Prompt Examples for our designed various levels of Reasoning Instructions.