Investigating Context Faithfulness in Large Language Models: The Roles of Memory Strength and Evidence Style

Yuepei Li, Kang Zhou, Qiao Qiao, Bach Nguyen, Qing Wang, Qi Li
Department of Computer Science
Iowa State University, Ames, Iowa, USA
{liyp0095, kangzhou, qqiao1, ntbach, qingwang, qli}@iastate.edu

Abstract

Retrieval-augmented generation (RAG) improves Large Language Models (LLMs) by incorporating external information into the response generation process. However, how context-faithful LLMs are and what factors influence LLMs' context faithfulness remain largely unexplored. In this study, we investigate the impact of memory strength and evidence presentation on LLMs' receptiveness to external evidence. We quantify the memory strength of LLMs by measuring the divergence in LLMs' responses to different paraphrases of the same question, which is not considered by previous works. We also generate evidence in various styles to examine LLMs' behavior. Our results show that for questions with high memory strength, LLMs are more likely to rely on internal memory. Furthermore, presenting paraphrased evidence significantly increases LLMs' receptiveness compared to simple repetition or adding details. These findings provide key insights for improving retrieval-augmented generation and context-aware LLMs. Our code is available at https://github.com/liyp0095/ContextFaithful.

1 Introduction

Retrieval-Augmented Generation (RAG) (Fan et al., 2024; Zhao et al., 2023) has gained increasing popularity as it improves the performance of Large Language Models (LLMs) by integrating external information during the generation process, particularly when the model's internal knowledge is insufficient or outdated (Bianchini et al., 2024; Procko, 2024; Siriwardhana et al., 2023; Vakayil et al., 2024; Wang et al., 2024; Jeong, 2023). It raises the importance of the study of how context-faithful LLMs are. In this study, we explore whether LLMs are context-faithful when encountering external information, particularly when that information conflicts with the LLMs' internal memory.

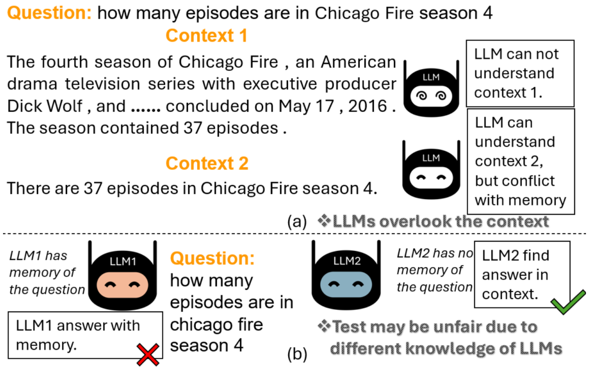

Question: how many episodes are in Chicago Fire season 4
Context 1
The fourth season of Chicago Fire , an American drama television series with executive producer Dick Wolf , and ...... concluded on May 17 , 2016 . The season contained 37 episodes .
LLM
LLM can not understand context 1.
Context 2
There are 37 episodes in Chicago Fire season 4.
LLM
LLM can understand context 2, but conflict with memory
(a) ❖LLMs overlook the context
LLM1 has memory of the question
LLM1
LLM1 answer with memory.
Question: how many episodes are in chicago fire season 4
LLM2
LLM2 has no memory of the question
LLM2 find answer in context.
(b) ❖Test may be unfair due to different knowledge of LLMs

Figure 1: Demonstration of issues in context faithfulness testing schema: a) LLMs may fail to comprehend long contexts, resulting in low receptiveness to the context. b) Evaluating LLMs on the same dataset may be unfair due to variations in their knowledge.

To investigate the issue of context faithfulness, there are two main approaches to creating knowledge conflict contexts. One approach (Longpre et al., 2021; Chen et al., 2022) is entity substitution, which replaces the gold entity in context with a similar one. Another approach (Xie et al., 2024; Jin et al., 2024) involves generating counter-memory evidence with LLMs, and these studies have shown that LLMs are generally receptive to external evidence as long as it is coherent.

These methods, however, overlook some important aspects of the task. First, previous work (Longpre et al., 2021; Xie et al., 2024) provides long contexts to LLMs, which can be challenging for LLMs to understand (Xie et al., 2024). It makes the test difficult to distinguish whether the cause of LLM behavior is due to knowledge conflicts or lacking comprehension ability. For example, in Figure 1 (a), the LLM may overlook both contexts, but the reason why Context 1 is ignored could be a lack of comprehension rather than knowledge conflicts. Second, different LLMs are trained with different data and are likely to obtain different knowledge. Thus, testing LLMs on the same dataset may be unfair. The test may favor LLMs with less knowledge. As shown in Figure 1 (b), LLMs with strong memory are less likely to be correct.

To address these issues, we introduce a method to quantify the memory strength of LLMs and generate short evidence in various styles to examine LLMs' behavior. Inspired by Zhao et al. (2024), we assess memory strength by measuring the divergence in LLMs' responses to different paraphrases of the same question. Intuitively, an LLM demonstrates high memory strength when it consistently provides the same answer across all paraphrased versions of a question. For evidence styles, we classify them into direct and indirect forms: direct evidence provides a straightforward answer to the question, while indirect evidence incorporates additional details to support the answer. Unlike Wan et al. (2024) that focuses on how different types of evidence influence LLMs' behavior, we study how the presentation of evidence affects its convincing availability. Through these methods, we analyze the relationship between context faithfulness and LLM memory strength, and we explore the impact of different evidence styles on context faithfulness. Our conclusions are as follows:

- The receptiveness of LLMs to external evidence is strongly correlated with memory strength to the question. We observed this relationship both across different datasets and different LLMs. Contrary to the findings of (Xie et al., 2024; Jin et al., 2024) that LLMs are highly receptive (less than 5%) to external evidence when it is coherent, we find that the probability of the model relying on its internal memory is non-negligible for questions that the LLMs have a strong memory. For example, GPT-4, which has strong memory on the NQ dataset, answers almost 50% of the questions with internal memory. We also demonstrate an urgent need for memory strength-aware evaluation metrics.

- The style of the evidence plays an important role in LLM's receptiveness to external information. Our research demonstrates that presenting the LLM with multiple paraphrases of the same evidence substantially increases its receptiveness. This approach outperforms simple repetition of the evidence and is more effective than adding additional details to the evidence. These findings provide valuable insights to the research of RAG.

2 Related Work

2.1 Context Faithfulness of LLM

To update static factual knowledge (Lazaridou et al., 2021; Karpukhin et al., 2020; Kasai et al., 2023) in LLMs, the retrieval-based method has been introduced to involve external information to LLMs (Lazaridou et al., 2022; Izacard et al., 2024; Khattab et al., 2022; Santhanam et al., 2022; Gao and Callan, 2022). However, these methods can introduce **knowledge conflicts** between the introduced external information (context) and pre-existing internal memories from LLMs. LLMs often rely on their internal memories, and overlook the contextual evidence (Longpre et al., 2021). To make LLMs more faithful to context, recent studies (Neeman et al., 2023; Li et al., 2023) fine-tune LLMs on counterfactual contexts, where the original facts are replaced with counterfactual ones. Another work (Zhou et al., 2023) proposes a novel approach using prompting to improve context faithfulness in LLMs without additional fine-tuning. Recent works (Farahani and Johansson, 2024; Wadhwa et al., 2024) found that models like T5, LLaMA2-7B, and Phi2-2.7B are generally context-faithful. However, they did not evaluate newer models (e.g., LLaMA3.2, GPT-4) or consider the impact of memory strength, limiting the generality of their conclusions.

A related area of research focuses on **prediction with abstention**. Neeman et al. (2023); Zhou et al. (2024) introduces answerability augmentation, where LLMs are trained to respond with "Unanswerable" when presented with irrelevant or randomly generated contexts. This ensures that the models do not make incorrect predictions without reliable evidence. Further studies (Wang et al., 2023, 2022) develop confidence calibration techniques to improve context faithfulness by encouraging LLMs to avoid overly confident predictions in ambiguous or uncertain situations.

In our work, we investigate the context faithfulness of LLMs when faced with conflicting knowledge. We define a model as context-faithful if it demonstrates high receptiveness to new facts and evidence with strong conflicting memories. This capability is essential to ensure the reliability of LLMs in the RAG system.

2.2 Construction of Knowledge Conflicts

In controlled experiments, knowledge conflicts are typically simulated by constructing counterfactual memories based on a model's internal memory. Var-

Origin Question	Paraphrased Questions	Answers with LLaMA2-7B
Who sings the Christmas song “Mary Did You Know”?	Can you tell me the artist who performs the Christmas song “Mary Did You Know”?	Of course! The artist who performs the Christmas song “Mary Did You Know” is Mark Lowry.
	Who is the singer behind the Christmas song “Mary Did You Know”?	The singer behind the Christmas song “Mary Did You Know” is Mark Lowry.
	Which musician sings the Christmas tune “Mary Did You Know”?	Brandy Clark
	Who is the vocalist that sings “Mary Did You Know” during the holiday season?	The vocalist who sings “Mary Did You Know” during the holiday season is Mark Lowry.

Table 1: LLMs give different answers when asked with paraphrased questions.

ious heuristic approaches have been proposed for this purpose, such as negation injection (Kassner et al., 2021; Petroni et al., 2020; Pan et al., 2021) and entity substitution (Longpre et al., 2021; Chen et al., 2022; Si et al., 2023; Zhou et al., 2023). Negation injection alters facts by introducing negations and entity substitution replaces mentions or entities in the evidence with alternatives to generate **counter-fact** evidence. However, these techniques are constrained to word-level edits, which can lead to low coherence across the constructed evidence. To address this limitation, recent studies (Xie et al., 2024; Jin et al., 2024) have explored generating evidence using LLMs, producing more coherent and consistent counterfactual content. We adopt this approach in generating our dataset, ensuring the generated evidence maintains a higher level of coherence.

3 Methodology

3.1 Problem Definition

Following prior work (Longpre et al., 2021; Chen et al., 2022; Xie et al., 2024), we adopt question answering (QA) task as the testbed for knowledge conflict experiments. For a given **question** Q, if the answer generated by the LLM relies solely on its internal parameters, it is referred to as the **memory answer** (MA). If an evidence passage E is provided with question Q, then ideally, LLM should generate an answer based on E, even if E conflicts with memory answers. We call the answers that conflict with MA as **counter memory answer** (CMA).

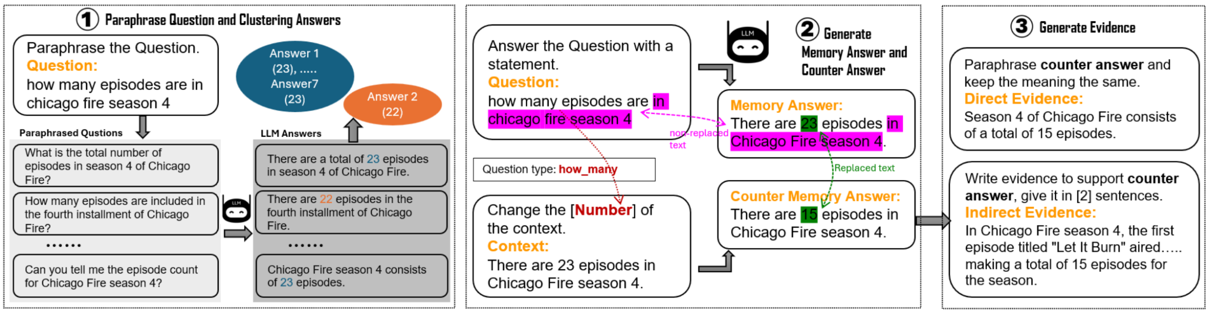
3.2 Datasets

We use two datasets for our experiments: the long-tail, entity-based QA dataset popQA, and the popular, human-written question dataset Natural Questions (NQ). Specifically:

- **popQA** (Mallen et al., 2023) is an entity-centric question-answering dataset comprising 14,000 questions. The dataset is derived from knowledge triples in Wikidata, where questions are generated using question templates specific to different relationship types. popQA aims to capture a realistic, long-tail distribution of entity popularity, making it a valuable resource for studying the performance of lesser-known entities. Xie et al. (2024) use popQA to test the receptiveness of LLMs by eliciting high-quality internal memory from LLMs and constructing the corresponding counter-memory. We reuse MA and CMA generated by Xie et al. (2024) for our experiments.
- **Natural Questions** (Kwiatkowski et al., 2019) is widely used in open-domain QA research. It consists of manually crafted questions based on selected paragraphs from Wikipedia, and the subjects in questions of the NQ dataset are generally more popular and commonly known. Longpre et al. (2021) provide a test set that is used to test the context faithfulness of LLMs by substituting entity of the NQ dataset. The entity substitute involves five categories: person (PER), date (DAT), numeric (NUM), organization (ORG), and location (LOC). The test set contains 4,685 samples, including 1,667 unique questions.

3.3 Memory Strength

Inspired by Zhao et al. (2024), we use the consistency of answers to different paraphrases of the same question Q to measure the LLM’s memory strength S_Q for the knowledge K_Q associated with the question. This method is motivated by the intuition that if an LLM does not have a strong memory of a question, it will often give different answers when asked with paraphrased questions that are semantically equivalent, as shown in Table 1. In contrast, it can produce consistent answers if the LLM has a strong memory of a question. The process involves two key steps: First, several para-


1 Paraphrase Question and Clustering Answers
Answer 1 (23), ... Answer7 (23)
Answer 2 (22)
Paraphrased Qustions
LLM Answers
LLM
2 Generate Memory Answer and Counter Answer
non-replaced text
Replaced text
3 Generate Evidence

Paraphrase the Question.
Question:
how many episodes are in chicago fire season 4

What is the total number of episodes in season 4 of Chicago Fire?

How many episodes are included in the fourth installment of Chicago Fire?

....

Can you tell me the episode count for Chicago Fire season 4?

There are a total of 23 episodes in season 4 of Chicago Fire.

There are 22 episodes in the fourth installment of Chicago Fire.

....

Chicago Fire season 4 consists of 23 episodes.

Answer the Question with a statement.
Question:
how many episodes are in chicago fire season 4

Question type: how_many

Change the [Number] of the context.
Context:
There are 23 episodes in Chicago Fire season 4.

Memory Answer:
There are 23 episodes in Chicago Fire season 4.

Counter Memory Answer:
There are 15 episodes in Chicago Fire season 4.

Paraphrase **counter answer** and keep the meaning the same.
Direct Evidence:
Season 4 of Chicago Fire consists of a total of 15 episodes.

Write evidence to support **counter answer**, give it in [2] sentences.
Indirect Evidence:
In Chicago Fire season 4, the first episode titled "Let It Burn" aired..... making a total of 15 episodes for the season.

(a) Memory Strength

(b) Memory / Counter Answer and Evidence Generation

Figure 2: Framework for Evaluating LLMs' Context Faithfulness. In Step 1, we calculate the memory strength of a question using the consistency of answers to different paraphrases of the same question. In Step 2, we generate MA (memory answer) under the closed-book setting and CMA (counter memory answer) by modifying answer entity in MA while keeping the other information. In Step 3, we generate supporting evidence for the CMA in various styles. In Step 4 (not shown), we test the LLM's response by presenting the questions with the evidence. All experiments are implemented under the zero-shot setting to avoid the bias introduced by demonstrations.

phrased versions of the original question are generated with ChatGPT[1], and the answers to those paraphrased questions are clustered (Section 3.3.1). Then, memory strength S_Q is calculated using answer consistency (Section 3.3.2).

3.3.1 Question Paraphrases and Answer Clustering

The prompt used for paraphrasing the question is provided in Tabel 11 (index 1) in the Appendix. For each question Q, we generate n paraphrases $\{P_1, \cdots, P_n\}_Q$. For the NQ dataset, we paraphrase the question in each data sample directly. For the popQA dataset, we paraphrase the question template for each relation type since all questions of the same relation type share the same question template. To ensure the paraphrased questions are proper to use, we check if two paraphrased questions are semantically equivalent with an LLM[2]. The prompt for this semantic equivalence detection is provided in Table 11 (index 2). For any paraphrase that is deemed not equivalent, we ask the LLM to re-generate it until a satisfactory version is produced.

Next, LLMs answer the paraphrased questions $\{P_1, \cdots, P_n\}_Q$ in a closed-book setting. We denote the answers as $\{A_1, \cdots, A_n\}_Q$. The answers are grouped into several clusters based on their consistency. The clustering is done by checking answers incrementally. If an answer matches any answer within an existing cluster, this answer is added to this cluster; if not, a new cluster is created with this answer. We use an LLM[2] to determine whether two answers are consistent. The prompt used for this answer inconsistency detection is shown in Table 11 (index 3). We denote the clusters for question Q as $\{c_1, \cdots, c_m\}_Q$.

3.3.2 Calculating Memory Strength

Once answer clusters $\{c_1, \cdots, c_m\}_Q$ are identified, memory strength $S(Q)$ can be obtained by calculating the negative entropy of cluster distribution. The formula is

$$S(Q) = \sum_{i=1}^{m} \frac{N(c_i)}{n} log \frac{N(c_i)}{n}, \tag{1}$$

where $N(c_i)$ is the number of answers in the cluster c_i, and n is the number of paraphrases for question Q. The memory strength score is a non-positive value. A larger score indicates a stronger memory (0 is the maximum value of memory strength score). In the experiments, we set $n = 7$ for all the questions in the NQ and popQA datasets. Memory strength reflects how well the LLM remembers the required knowledge: the weaker the memory, the more random and inconsistent the answers are.

3.4 MA, CMA, and Evidence Generation

3.4.1 MA and CMA Generation

For the popQA dataset, both MA and CMA are obtained following the method described in Xie et al. (2024). For the MA of the NQ dataset, we also use a closed-book approach, similar to Xie et al. (2024). While, the process of generating CMA differs from the process of generating CMA in Xie et al. (2024). Unlike the popQA dataset, the NQ dataset does not

[1] https://platform.openai.com/docs/models/gpt-3-5-turbo, the specific version is 0125.

[2] https://huggingface.co/meta-llama/Meta-Llama-3.1-8B

provide relation types for the questions or offer sets of subject and object entities for substitution. To address this issue, we propose an approach using an LLM to substitute entities in MA to generate CMA. First, we identify which "wh-" question type[3] the question belongs to using string matching. Then, based on the question type, we determine the type of entity to be replaced in the MA. Finally, we use an LLM to make the substitution. For example, in Figure 2, the question "how many episodes..." is of the type "how_many", so the entity to be replaced in the MA "there are 23 episodes..." should be a *NUMBER*. We let ChatGPT perform the substitution with an alternative entity. The prompt used is shown in Table 11 (index 5). We have the detailed description for generating CMA in Appendix A.1.

CMA filter. As noted in Section 3.3, LLMs can produce multiple MAs for the same question. To ensure the CMA conflicts with MAs, we require that the CMA is different from any of the answers $\{A_1, \cdots, A_n\}$ generated in Section 3.3.1, so the alternative entity should not appear in MAs. For the popQA dataset, the alternative entity is known. For the NQ dataset, we first identify the alternative entity in the CMA by comparing the MA and CMA, and then check if this entity appears in any of the MAs $\{A_1, \cdots, A_n\}$. We filter out data samples whose CMA does not conflict with MAs.

3.4.2 Evidence Generation

In this section, we explain how to generate different styles of evidence. We classify evidence into two categories: direct evidence and indirect evidence.

Direct evidence is a semantically equivalent statement of the CMA, providing the clearest support for the claim made by the CMA. We generate the direct evidence by paraphrasing the CMA with ChatGPT, following the prompt shown in Table 11 (index 6). For example, in Figure 2, the CMA "there are 15 episodes in Chicago Fire season 4" is paraphrased to "season 4 of Chicago Fire consists of a total of 15 episodes". These two statements are semantically equivalent.

To ensure the reliability of the evidence, the evidence must mutually entail with the CMA. This entailment is verified using an NLI model [4]. Direct evidence is intuitive, simple, and coherent, making it the straightforward type of evidence for the LLM to process. If the LLM is receptive to external evidence, it should be able to adopt direct evidence easily.

Indirect evidence differs from direct evidence by adding extra details that provide a more thorough description of the subject related to the CMA. This additional information makes the evidence more comprehensive and might be more persuasive. For example, in Figure 2, the indirect evidence includes details not found in the counter answer, such as the title of the first episode and its release date, along with the fact that there are 15 episodes in total. The prompt to generate indirect evidence is shown in Table 11 (index 7).

To ensure the reliability of indirect evidence, the indirect evidence should entail the CMA and the additional information introduced by the evidence should not entail the MA. Otherwise, the indirect evidence can support both the MA and CMA. The NLI model[4] is used to verify that indirect evidence entailed with CMA and neutral or contradictory with MA.

For both direct and indirect evidence, if the content generated by the LLM does not meet the required conditions, we prompt the LLM to regenerate it up to five times. If it still fails after five attempts, we exclude that question from the dataset.

4 Experiments

In this study, we aim to investigate two key research questions. 1) Does memory strength have an impact on the context faithfulness of LLMs? 2) Does the style of evidence affect the context faithfulness of LLMs? These research questions are explored in Section 4.2 and 4.3, respectively. We also provide additional studies in Appendix B, which includes a study about the impact of option order and a case study.

4.1 Experiment Setup

LLM Models. Our experiments are conducted using six well-known language models: ChatGPT (OpenAI, 2023a), GPT-4 (OpenAI, 2023b), LLaMA2-7B, LLaMA2-70B (Touvron et al., 2023), LLaMA3.2-3B (Meta, 2024), and Claude3.5 (Anthropic, 2024). These models represent a diverse range of architectures and capabilities. ChatGPT, GPT-4, and Claude3.5 are cutting-edge models developed by OpenAI and Anthropic. LLaMA2, with its 7 billion and 70 billion parameter variants, is a strong open-source alternative that has demon-

[3] which refers to what, when, where, who, whom, which, whose, why, and how.

[4] https://huggingface.co/microsoft/deberta-v2-xxlarge-mnli

strated competitive performance across a wide variety of tasks. LLaMA3.2-3B is the newest version of LLaMA. The inclusion of models with varying scales (from 3B to 70B) and training methodologies allows us to explore both closed-source systems (GPT-4, ChatGPT, and Claude3.5) and open-source solutions (LLaMA2-7B, LLaMA-70B and LLaMA3.2-3B).

Evaluation Metrics. Following previous work (Longpre et al., 2021; Xie et al., 2024; Chen et al., 2022), we transform the short answer QA to a multiple-choice QA format by providing a few options as possible answers[5]. This limits the answer generation space and makes it easy to evaluate without manual checking. Specifically, for each question from both datasets, LLMs are instructed to select one answer from the MA, CMA, and "Uncertain" (UCT). We report the ratio of MA (R_m), CMA (R_c), and UCT (R_u) as calculated below:

$$\begin{aligned} R_m &= \frac{f_m}{f_m + f_c + f_u} \\ R_c &= \frac{f_c}{f_m + f_c + f_u} \\ R_u &= \frac{f_u}{f_m + f_c + f_u}, \end{aligned} \tag{2}$$

where f_m, f_c, and f_u are the count of questions with MA, CMA, and UCT answers, respectively.

4.2 Role of Memory Strength

4.2.1 Correlation between Context Faithfulness and Memory Strength

To demonstrate the relationship between context faithfulness and memory strength, we categorize the questions in each dataset into four groups according to the memory strength of each LLM. The four groups are low, mid-low, mid-high, and high, corresponding to the memory strength intervals [-2, -1], (-1, -0.5], (-0.5, -0.25] and [-0.25, 0], respectively. We use the direct evidence in this experiment. The results are shown in Figure 3[6]. Figure 3 (a)(b) shows the ratios of questions with MA, CMA, and UCT answers for the NQ and popQA datasets, respectively. Note that different LLMs may have different memory strengths to the same question. Therefore, both the specific questions and the count of questions in the same group can vary across different LLMs. To illustrate this, we present the count of questions in each group (low, mid-low, mid-high, and high) in Figure 3 (c)(d) for popQA and NQ datasets, respectively. We can draw the following conclusions.

There is a clear positive correlation between memory strength and MA ratio for individual LLMs. From Figure 3 (a)(b), it is obvious that for all tested LLMs, the ratio of MA (red) increases when memory strength increases, while the ratio of CMA (blue) decreases. This trend is also consistent across both datasets. It is more obvious for GPT models. Among the tested LLMs, Claude3.5 tends to choose UCT options more often, especially on questions with high memory strength.

Memory Strength Increases with Model Scale. We can observe from Figure 3 (c)(d) that larger LLMs, such as GPT-4, have more data samples in the high memory strength group (hi), while smaller LLMs, such as LLaMA3.2-3B, have more samples in the low memory strength group (lo). This aligns with the common intuition that larger LLMs, with more parameters, have more knowledge than smaller LLMs. Further evidence and discussion can be found in the Appendix B.1.

4.2.2 LLMs Performance Analysis

In the aforementioned conclusion, different LLMs have different knowledge. Thus, testing LLMs' context faithfulness on the same dataset may be unfair. To illustrate this, we compute the average memory strength of LLMs on PopQA and NQ datasets, respectively, along with their MA and UCT ratios. The results are presented in Figure 4. We can draw the following conclusions.

First, for different LLMs, a lower average memory strength does not necessarily imply better context faithfulness. For example, Claude 3.5 has a high average memory strength, implying that Claude 3.5 is a knowledgeable LLM, but it has a low MA ratio (R_m) and a high UCT ratio (R_u). This indicates that Claude 3.5 tends to refuse to answer when facing knowledge conflicts. In contrast, LLaMA3.2-3B, despite having much less knowledge (low average memory strength), has the second-highest MA ratio (R_m). This means that LLaMA3.2-3B relies heavily on limited internal knowledge when facing knowledge conflicts, implying that it is not context-faithful.

Second, newer versions of GPT and LLaMA models appear to overlook context faithfulness issues during the training process. GPT-4, with

[5] Xie et al. (2024) shows that answer consistency between short answer and multi-choice are 94%, 96% and 92% for ChatGPT, GPT-4 and LLaMA2-7B, respectively.

[6] We put results for other evidence styles in Appendix, Figures 8, 9, and 10. The conclusion is consistent.

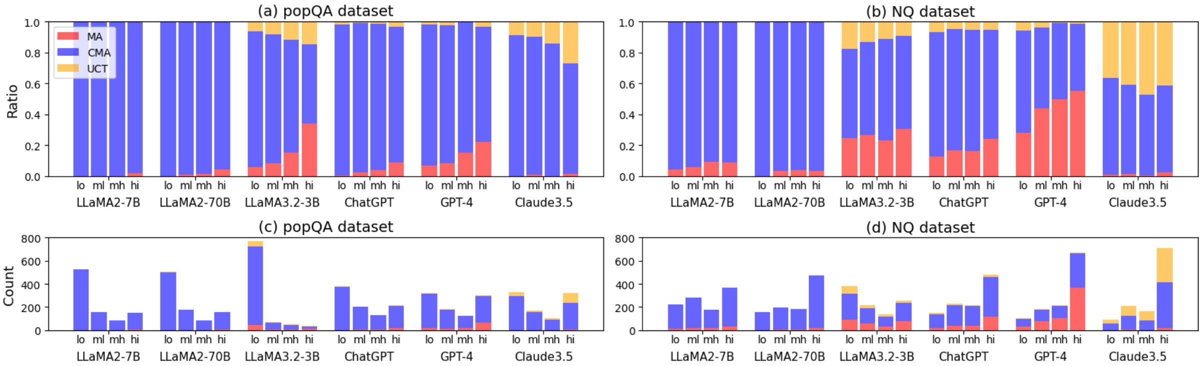


Figure 3: Relationship between Memory Strength and Different Answers' Ratios across popQA and NQ Datasets (with Direct Evidence). The figure presents the ratio and count of MA, CMA, and UCT across four memory strength groups: low(lo), mid-low(ml), mid-high(mh), and high(hi). The results show a clear positive correlation between memory strength and MA ratio(R_m).

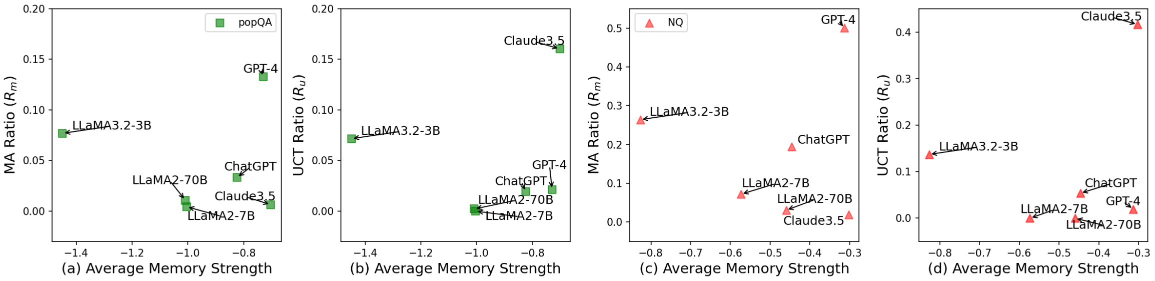


Figure 4: MA Ratio (R_m) and UCT Ratio (R_u) VS Memory Strength across PopQA and NQ Datasets. The figure presents the relationship between average memory strength and the MA ratio as well as the UCT ratio for all six LLMs. It shows that lower average memory strength does not always mean better context faithfulness for different LLMs and newer versions of GPT and LLaMA models seem to ignore context faithfulness issues.

slightly more knowledge than Chatgpt, shows a significantly higher MA ratio (R_m), and LLaMA3.2-3B, with the least knowledge, shows a higher MA ratio (R_m) than the two LLaMA2 models.

Finally, a new context faithfulness evaluation metric is needed. This metric should consider both the different answers' ratios and memory strength when LLMs encounter knowledge conflicts. Simply using the MA ratio (which is used widely to measure context faithfulness currently) to evaluate context faithfulness across different LLMs may not be fair and sufficient.

4.3 Role of Evidence Style

Evidence Styles. We formulate four types of evidence styles: **1)** Direct Evidence. This is the most straightforward form of evidence and serves as our baseline. To assess the impact of evidence length, we also create versions where the direct evidence is repeated twice and three times for comparison. **2)** Direct Evidence Combined with Paraphrases of CMA. To examine the effect of evidence phrasing and expression, we combine the direct evidence with one paraphrase of the CMA to form a two-sentence evidence and with two paraphrases to form a three-sentence evidence. **3)** Indirect Evidence. We generate indirect evidence consisting of two sentences and three sentences, respectively[7]. **4)** Direct Evidence Combined with Indirect Evidence. We combine the direct evidence with the first sentence of the two-sentence indirect evidence to form a two-sentence evidence and with both sentences to form a three-sentence evidence.

Table 2 presents the final number of instances used for evaluation. We observe a slight difference in the quantities of questions with direct and indirect evidence since it is easier for ChatGPT to generate direct evidence that meets our requirements. The specific number of instances at each step in evidence generation is detailed in Table 7 in the Appendix. Due to the quantity difference between direct evidence and indirect evidence, we divide the styles of evidence into two groups: Group 1

[7] We regulate the length of the generated evidence to control the influence of evidence length. The prompts used are detailed in Table 11 (index 7) in Appendix.

	popQA						NQ					
	LLaMA2-7B	**LLaMA2-70B**	**LLaMA3.2-3B**	**ChatGPT**	**GPT-4**	**Claude3.5**	**LLaMA2-7B**	**LLaMA2-70B**	**LLaMA3.2-3B**	**ChatGPT**	**GPT-4**	**Claude3.5**
# of Q (Initial)	1000	1000	1000	1000	1000	1000	1667	1667	1667	1667	1667	1667
# of Q with direct evidence	918	922	938	933	933	931	1042	1009	1002	1079	1171	1173
# of Q with indirect evidence	901	895	917	911	918	913	976	972	941	1025	1108	1059

Table 2: Number of final examples for each LLM. The difference between LLMs is due to their different outputs going through the framework.

Dataset	Evidence Style	S #	LLaMA2-7B			LLaMA2-70B			LLaMA3.2-3B			ChatGPT			GPT-4			Claude3.5		
			$R_m \downarrow$	$R_c \uparrow$	R_u	$R_m \downarrow$	$R_c \uparrow$	R_u	$R_m \downarrow$	$R_c \uparrow$	R_u	$R_m \downarrow$	$R_c \uparrow$	R_u	$R_m \downarrow$	$R_c \uparrow$	R_u	$R_m \downarrow$	$R_c \uparrow$	R_u
popQA	Group 1: Q with direct evidence																			
	Direct Evidence	1	0.44	99.56	0.0	1.08	98.7	0.22	7.69	85.15	7.16	3.32	94.75	1.93	13.29	84.57	2.14	0.65	83.31	16.04
		2	0.44	99.56	0.0	0.98	98.81	0.22	2.67	95.83	1.5	2.79	96.03	1.18	4.39	93.46	2.14	0.32	94.08	5.6
		3	0.65	99.35	0.0	1.08	98.7	0.22	2.56	94.98	2.46	3.0	95.5	1.5	2.57	95.28	2.14	0.21	94.73	5.06
	Direct+Paraphrase	2	0.22	99.78	0.0	1.19	98.7	0.11	3.1	95.62	1.28	2.36	97.0	0.64	3.0	95.71	1.29	0.21	95.69	4.09
		3	0.11	99.89	0.0	0.43	99.35	0.22	**1.07**	97.86	1.07	**1.39**	**98.28**	0.32	**1.29**	**98.5**	0.21	**0.11**	**98.06**	1.83
	Group 2: Q with indirect evidence																			
	Direct Evidence	1	0.44	99.56	0.0	0.45	99.33	0.22	7.31	85.61	7.09	3.07	95.28	1.65	12.53	85.29	2.18	0.66	83.9	15.44
	Indirect Evidence	2	0.0	100.0	0.0	0.11	99.89	0.0	4.36	89.64	6.0	3.18	96.38	0.44	13.51	85.73	0.76	0.88	82.37	16.76
		3	**0.0**	**100.0**	0.0	**0.0**	**100.0**	0.0	3.27	92.58	4.14	1.76	97.91	0.33	9.26	90.2	0.55	0.88	87.95	11.17
	Direct+Indirect	2	0.22	99.78	0.0	0.11	99.78	0.11	3.71	93.24	3.05	1.87	97.26	0.88	7.95	91.18	0.87	0.88	93.43	5.7
		3	0.11	99.89	0.0	0.11	99.78	0.11	1.42	**98.15**	0.44	1.43	98.13	0.44	5.12	94.77	0.11	0.88	96.28	2.85
NQ	Group 1: Q with direct evidence																			
	Direct Evidence	1	7.2	92.8	0.0	3.07	96.93	0.0	26.41	59.88	13.71	19.46	75.16	5.38	50.04	47.99	1.96	1.96	56.4	41.64
		2	5.47	94.53	0.0	3.07	96.93	0.0	16.93	75.2	7.86	19.09	76.83	4.08	20.24	77.54	2.22	0.77	83.7	15.53
		3	6.81	93.19	0.0	2.68	97.22	0.1	11.79	76.11	12.1	22.06	72.75	5.19	17.34	80.87	1.79	**0.26**	88.22	11.52
	Direct+Paraphrase	2	4.13	95.87	0.0	1.49	98.41	0.1	13.21	79.33	7.46	15.29	81.28	3.43	18.96	79.67	1.37	0.34	88.65	11.01
		3	**3.26**	**96.74**	0.0	**1.19**	**98.61**	0.2	9.38	83.27	7.36	9.55	86.75	3.71	**11.27**	**87.28**	1.45	**0.26**	**93.09**	6.65
	Group 2: Q with indirect evidence																			
	Direct Evidence	1	5.53	94.47	0.0	2.67	97.32	0.0	23.19	63.73	13.08	18.73	75.71	5.56	48.83	49.19	1.99	1.77	59.44	38.79
	Indirect Evidence	2	3.28	95.29	1.43	1.65	98.25	0.1	13.32	77.77	8.92	13.66	84.1	2.24	44.59	53.7	1.71	1.98	67.88	30.14
		3	4.82	94.06	1.13	1.85	97.84	0.31	11.53	80.86	7.61	13.27	84.19	2.54	39.89	58.57	1.53	3.65	71.32	25.03
	Direct+Indirect	2	5.33	94.67	0.0	1.34	98.25	0.41	9.75	82.28	7.97	12.68	84.78	2.54	32.4	66.06	1.53	1.36	80.92	17.73
		3	4.41	95.59	0.0	1.44	98.56	0.0	**8.32**	**85.49**	6.18	**9.46**	**88.1**	2.44	28.7	69.67	1.62	1.77	83.94	14.29

Table 3: Results of LLM Receptiveness to Different Evidence Styles Across NQ and popQA Datasets. The table presents the MA ratio (R_m), CMA ratio (R_c), and uncertain answer ratio (R_u) for various evidence styles across six models. All the ratios are in %. The best results are highlighted in **bold**, and the second-best results are underlined.

includes Direct Evidence and Direct + Paraphrase evidence. Group 2 includes Indirect Evidence and Direct + Indirect evidence. Each group has different Direct Evidence results serving as baselines.

Table 3 shows the results of different evidence styles. We can make the following observations and conclusions.

In Group 2, the MA Ratio (R_m) of direct evidence is slightly lower than that in Group 1. During the evidence generation, there are some questions for which ChatGPT can provide direct evidence but cannot produce indirect evidence. Removing these questions leads to a decrease in R_m of direct evidence with one sentence, which implies that LLMs have a relatively high R_m for the removed questions. But in general, the R_m of direct evidence with one sentence in Group 1 is close to that in Group 2, so the results of Group 1 and Group 2 are **still comparable**.

Simple repetition of direct evidence is not always effective. Comparing direct evidence with one to three sentences, we observe similar R_m and R_c for LLaMA2-7B, LLaMA2-70B, and ChatGPT. For LLaMA3.2-3B, GPT-4, and Claude3.5, R_m of direct evidence with two and three sentences decreases significantly. The results imply that the newer LLMs are receptive to evidence repeated multiple times.

Paraphrasing direct evidence is highly effective across all models and datasets. Comparing Direct Evidence with two and three sentences and Direct + Paraphrase with two and three sentences, respectively, we observe R_m of the latter significantly decreases. For example, R_m is reduced by more than half, comparing Direct + Paraphrase with three sentences with Direct Evidence with three sentences on the popQA dataset for all tested LLMs. The result implies that paraphrasing is an effective method to enhance the receptiveness of LLMs to external evidence.

Indirect Evidence improves LLMs' receptiveness to CMAs, but less effectively than paraphrasing. Comparing Indirect Evidence with two and three sentences with Direct Evidence with one sentence, R_m decreases for almost all LLMs, but the reduction is not significant compared to the Direct + Paraphrase evidence with two or three sentences. It implies that adding detailed information is less effective than paraphrasing direct evidence.

Combining Direct evidence with Indirect evidence generally enhances persuasiveness. Comparing Direct + Indirect evidence with Indirect Evidence, R_m decreases except for LLaMA2-7B. For example, comparing Direct + Indirect with three sentences and Indirect Evidence with three sentences, R_m has an obvious decrease. The result implies that adding direct evidence to indirect evidence is effective in improving LLMs' receptiveness to CMAs.

5 Conclusion

We investigate how context-faithful LLMs are to external evidence across two datasets, PopQA and NQ datasets, using LLaMA2-7B, LLaMA2-70B, ChatGPT, GPT-4, LLaMA3.2-3B and Claude3.5. Our findings highlight the critical role of memory strength in shaping LLM behavior. There is a clear positive correlation between memory strength and memory answer ratio. Furthermore, we demonstrate that paraphrasing significantly enhances the context faithfulness of LLMs across various models and datasets. These findings offer valuable insights for advancing research in retrieval-augmented generation and context-based LLM applications.

Limitations

Our framework does not process all types of questions in the NQ dataset. Although it effectively handles the majority of NQ questions, it currently lacks the capability to address "what," "how," and "why" question types. The omission of these questions may introduce some bias into our results. Similar to previous studies, our study also focuses on knowledge conflict for extractive QA tasks, where the answer must appear in the evidence. Our conclusion may not be extendable to other types of QA tasks, such as abstractive QA and generative QA.

We employed a Natural Language Inference (NLI) model to detect and filter the generated data. Although the NLI model demonstrates high accuracy and the quality of generated data is high, it still cannot guarantee complete correctness. Further, since the NLI model is also trained using language models, which may be biased with parametric memory, it may introduce biases facing knowledge conflicts.

Acknowledgments

The work was supported in part by the US National Science Foundation under grant NSF-CAREER 2237831. We also want to thank the anonymous reviewers for their helpful comments.

References

Anthropic. 2024. Claude 3.5. https://www.anthropic.com/claude. Accessed: 2024-12-15.

Filippo Bianchini, Marco Calamo, Francesca De Luzi, Mattia Macrì, and Massimo Mecella. 2024. Enhancing complex linguistic tasks resolution through fine-tuning llms, rag and knowledge graphs (short paper). In *International Conference on Advanced Information Systems Engineering*, pages 147–155. Springer.

Anthony Chen, Pallavi Gudipati, Shayne Longpre, Xiao Ling, and Sameer Singh. 2021. Evaluating entity disambiguation and the role of popularity in retrieval-based NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4472–4485, Online. Association for Computational Linguistics.

Hung-Ting Chen, Michael Zhang, and Eunsol Choi. 2022. 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*, pages 2292–2307, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6491–6501.

Mehrdad Farahani and Richard Johansson. 2024. Deciphering the interplay of parametric and non-parametric memory in retrieval-augmented language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16966–16977.

Luyu Gao and Jamie Callan. 2022. Unsupervised corpus aware language model pre-training for dense passage retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2843–2853, Dublin, Ireland. Association for Computational Linguistics.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2024. Atlas: few-shot learning with retrieval augmented language models. *J. Mach. Learn. Res.*, 24(1).

Cheonsu Jeong. 2023. Generative ai service implementation using llm application architecture: based on rag model and langchain framework. *Journal of Intelligence and Information Systems*, 29(4):129–164.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Li Qiuxia, and Jun Zhao. 2024. Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16867–16878, Torino, Italia. ELRA and ICCL.

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

Jungo Kasai, Keisuke Sakaguchi, yoichi takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, and Kentaro Inui. 2023. Realtime qa: What's the answer right now? In *Advances in Neural Information Processing Systems*, volume 36, pages 49025–49043. Curran Associates, Inc.

Nora Kassner, Oyvind Tafjord, Hinrich Schütze, and Peter Clark. 2021. BeliefBank: Adding memory to a pre-trained language model for a systematic notion of belief. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8849–8861, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp. *arXiv preprint arXiv:2212.14024*.

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

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

Angeliki Lazaridou, Adhi Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, et al. 2021. Mind the gap: Assessing temporal generalization in neural language models. *Advances in Neural Information Processing Systems*, 34:29348–29363.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2023. Large language models with controllable working memory. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1774–1793, Toronto, Canada. Association for Computational Linguistics.

Yuepei Li, Kang Zhou, Qiao Qiao, Qing Wang, and Qi Li. 2025. Re-examine distantly supervised NER: A new benchmark and a simple approach. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10940–10959, Abu Dhabi, UAE. 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. *arXiv preprint arXiv:2109.05052*.

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. Llama 3.2: Open and efficient large language models. https://ai.facebook.com/llama. Accessed: 2024-12-15.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2023. DisentQA: Disentangling parametric and contextual knowledge with counterfactual question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10056–10070, Toronto, Canada. Association for Computational Linguistics.

OpenAI. 2023a. Chatgpt: Large Language Model. https://chat.openai.com. September 2023 version.

OpenAI. 2023b. Gpt-4 technical report. https://openai.com/research/gpt-4. March 2023 version.

Liangming Pan, Wenhu Chen, Min-Yen Kan, and William Yang Wang. 2021. Contraqa: Question answering under contradicting contexts.

Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2020. How context affects language models' factual predictions. In *Automated Knowledge Base Construction*.

Tyler Procko. 2024. Graph retrieval-augmented generation for large language models: A survey. *Available at SSRN*.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734, Seattle, United States. Association for Computational Linguistics.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Lee Boyd-Graber, and Lijuan Wang. 2023. Prompting GPT-3 to be reliable. In *The Eleventh International Conference on Learning Representations*.

Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana, and Suranga Nanayakkara. 2023. Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering. *Transactions of the Association for Computational Linguistics*, 11:1–17.

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

Sonia Vakayil, D. Sujitha Juliet, Anitha. J, and Sunil Vakayil. 2024. Rag-based llm chatbot using llama-2. In *2024 7th International Conference on Devices, Circuits and Systems (ICDCS)*, pages 1–5.

Hitesh Wadhwa, Rahul Seetharaman, Somya Aggarwal, Reshmi Ghosh, Samyadeep Basu, Soundararajan Srinivasan, Wenlong Zhao, Shreyas Chaudhari, and Ehsan Aghazadeh. 2024. From rags to rich parameters: Probing how language models utilize external knowledge over parametric information for factual queries. *CoRR*.

Alexander Wan, Eric Wallace, and Dan Klein. 2024. What evidence do language models find convincing? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7468–7484, Bangkok, Thailand. Association for Computational Linguistics.

Chengrui Wang, Qingqing Long, Xiao Meng, Xunxin Cai, Chengjun Wu, Zhen Meng, Xuezhi Wang, and Yuanchun Zhou. 2024. Biorag: A rag-llm framework for biological question reasoning. *arXiv preprint arXiv:2408.01107*.

Haoyu Wang, Hongming Zhang, Yuqian Deng, Jacob Gardner, Dan Roth, and Muhao Chen. 2023. Extracting or guessing? improving faithfulness of event temporal relation extraction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 541–553, Dubrovnik, Croatia. Association for Computational Linguistics.

Yiwei Wang, Muhao Chen, Wenxuan Zhou, Yujun Cai, Yuxuan Liang, Dayiheng Liu, Baosong Yang, Juncheng Liu, and Bryan Hooi. 2022. Should we rely on entity mentions for relation extraction? debiasing relation extraction with counterfactual analysis. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3071–3081, Seattle, United States. Association for Computational Linguistics.

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*.

Ruochen Zhao, Hailin Chen, Weishi Wang, Fangkai Jiao, Xuan Long Do, Chengwei Qin, Bosheng Ding, Xiaobao Guo, Minzhi Li, Xingxuan Li, et al. 2023. Retrieving multimodal information for augmented generation: A survey. *arXiv preprint arXiv:2303.10868*.

Yukun Zhao, Lingyong Yan, Weiwei Sun, Guoliang Xing, Chong Meng, Shuaiqiang Wang, Zhicong Cheng, Zhaochun Ren, and Dawei Yin. 2024. Knowing what LLMs DO NOT know: A simple yet effective self-detection method. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7051–7063, Mexico City, Mexico. Association for Computational Linguistics.

Kang Zhou, Yuepei Li, and Qi Li. 2022. Distantly supervised named entity recognition via confidence-based multi-class positive and unlabeled learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7198–7211, Dublin, Ireland. Association for Computational Linguistics.

Kang Zhou, Yuepei Li, Qing Wang, Qiao Qiao, and Qi Li. 2024. GenDecider: Integrating "none of the candidates" judgments in zero-shot entity linking reranking. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 239–245, Mexico City, Mexico. Association for Computational Linguistics.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14544–14556, Singapore. Association for Computational Linguistics.

Appendix

Within this supplementary material, we elaborate on the following aspects:

- A Methodology Details
- B Additional Studies
- C Prompts

A Methodology and Experiment Details

A.1 CMA Generation for NQ dataset

Departing from prior approaches (Longpre et al., 2021) that use NER techniques (Chen et al., 2021; Zhou et al., 2022; Li et al., 2025) to locate and substitute answer entities, we employ LLMs to automatically generate alternative entities. Our CMA construction process from MA consists of three steps: 1) identify question type, 2) determine entity type in MA to change, and 3) generate CMA with LLMs.

Identity Question Type: We first build a typing tree using rules to categorize questions. Figure 5 illustrates the typing tree, which consists of a two-layer structure. In the typing process, we first determine if a question begins with one of the following words: “what”, “when”, “where”, “which”, “who”, “why”, or “how”. However, this approach can still group different types of questions together. To address this, we use a second layer to refine the typing by analyzing two specific words in the question. For example, the question shown in Figure 2 falls into the “how_many” category. Table 4 shows the statistics of question types of the NQ dataset. Note that, we find 127 samples that are not questions in the process, so we list them as “other”.

Determine Entity Type in MA to Change: After categorizing the questions, we determine the entity type in MA needs to be replaced. To achieve this, we give each type of question an entity type, and many questions can share the same entity type. For example, both “when” and “what year” ask for a time. So a time entity in MA should be substituted. The final set of entity types is summarized in Table 5. We do not process questions starting with “what”, “which” or “how” due to the lack of a unified entity type for these questions. Table 6 shows the statistics of the unprocessed questions. These question types account for only 7.3% of the NQ dataset (with 2, 98, 22, and 0 instances for “how”, “what”, “which”, and “why”, respectively), and this issue does not occur in the PopQA dataset.

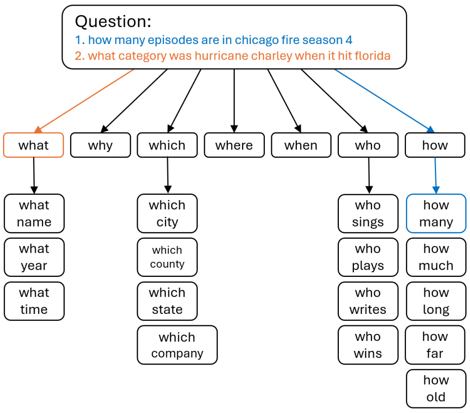


Figure 5: Two-layer Question Typing Tree

Question Type	Count
how_many	97
how_much	1
how_long	3
how_old	3
how	2
who_sings	100
who_plays	179
who_writes	65
who_wins	55
who	479
where	138
when	276
what_year	7
what_name	4
what	98
which_country	6
which_city	2
which_state	2
which_year	1
which	22
why	0
other	127
total	1667

Table 4: Distribution of Question Types and their Counts

Since the overall trends are consistent across both datasets, we believe the missing questions will not affect our conclusions.

Generate CMA with LLMs: Instead of manually editing the MA, we leverage the LLMs’ ability to generate CMA by providing it with a carefully designed prompt, which is shown in Table 11 (index 5). This prompt instructs the LLM to replace the entity with a certain type in the MA (from Step 2) with an alternative, ensuring the generated CMA differs from the MA.

The generated CMA must meet two key criteria: 1) The CMA must directly contradict the MA. To ensure this, we employ a Natural Language Inference (NLI) model[8] to verify the contradiction between the two answers. 2) The alternative entity in CMA must not appear in the question. We achieve

[8] https://huggingface.co/microsoft/deberta-v2-xxlarge-mnli

Question Type	Key Term
when, what_year, which_year, how_long	time
where, which_city, which_state, which_country	location
who, what_name	name of person
how_many, how_much	number
who_sings	singer's name
who_plays	player's name
who_writes	writer's name
who_wins	winner's name
how_far	distance
how_old	age

Table 5: Question Types and Their Corresponding Key Terms

Type	Count	Question Examples
how	2	how are leaders of the two parties in congress chosen
what	98	what is the setting of the story sorry wrong number
which	22	which domain of life are humans members of
why	0	-
other	127	latest season on keeping up with the kardashians
total	1667	-

Table 6: Summary of Excluded Question Types in Memory Answer and Counter Answer Generation. The table lists question types that were excluded from processing due to either the difficulty in identifying a unified entity type ("how", "what", "which") or not question ("other").

this check by string matching. If the CMA fails to meet either of these criteria, we prompt the LLM to regenerate the CMA up to 5 times. If no proper CMA is generated, we filter out this question.

A.2 Dataset Details

For the popQA dataset, we use the dataset from Xie et al. (2024) by randomly selecting 1,000 questions from the data intersection of the conflicts generated by LLaMA2-7B, LLaMA2-70B, LLaMA3.2-3B, ChatGPT, GPT-4 and Claude3.5. We use the MA and CMA from Xie et al. (2024) and only generate direct evidence and indirect evidence using our framework. For the NQ dataset, we use the test set from Longpre et al. (2021), which consists of 1,667 unique questions. The MA, CMA, and evidence are all generated with our framework. The dataset scale at each step is presented in Table 7.

A.3 Human Evaluation for Model Reliability

To ensure the reliability of the NLI model, Xie et al. (2024) randomly sample 200 generated examples and manually annotate whether the generated content entails the corresponding claim. The labels are supportive (entailment in the NLI task) or not supportive (either neutral or contradiction in the NLI task). The accuracy is 99%.

Following this process, we evaluate how reliable the generated CMA is. We randomly sample 200 generated examples in the NQ dataset and manually annotate whether the correct entity in MA is found and replaced with a same type alternative. The accuracy is 98%, which means the generated CMA is reliable.

B Additional Studies

B.1 Memory Strength on Different Datasets

We illustrate the distributions of memory strength on the popQA and NQ datasets for LLaMA2-7B, LLaMA2-70B, LLaMA3.2-3B, ChatGPT, GPT-4 and Claude3.5, respectively (shown in Figure 6). The results show that **LLMs demonstrate stronger memory for the NQ dataset than the popQA dataset.** For the NQ dataset, most questions fall within the bin of (0.25, 0]. Only a few questions fall within bins of weaker memory strength. In contrast, the popQA dataset has a greater number of questions in bins with weaker memory strength. This phenomenon is consistent across all six evaluated LLMs, indicating that the LLMs have more knowledge of the NQ dataset compared to the PopQA dataset. A possible explanation is that the NQ dataset covers more commonly discussed subjects than those in the PopQA dataset. These subjects may have been encountered more frequently during the training of the LLMs, making it easier for the models to recall the information and resulting in stronger memory strength.

B.2 Order of Options

To test the effect of the order of options on R_m, we conduct an experiment with one sentence direct evidence by changing the order of options (MA option and CMA option). We define the scenario where the CMA option is presented first in the prompt as "CMA first", and the scenario where the MA option is presented first as "MA first"[9]. Figure 7 shows the results.

Across all six models (LLaMA2-7B, LLaMA2-70B, LLaMA3.2-3B, ChatGPT, GPT-4 and Claude3.5), we observe a consistent trend: MA ratio (R_m) under "CMA first" is significantly higher than that under "MA first". Evaluations under "CMA first" demonstrate that LLMs are less context-faithful.

To further demonstrate the effect of the order of options on R_m, we compare the performance of experiments with "CMA first" and "MA first"

[9] All previous evaluations are under "MA first" conditions.

	popQA						NQ					
	LLaMA2-7B	**LLaMA2-70B**	**LLaMA3.2-3B**	**ChatGPT**	**GPT-4**	**Claude3.5**	**LLaMA2-7B**	**LLaMA2-70B**	**LLaMA3.2-3B**	**ChatGPT**	**GPT-4**	**Claude3.5**
Initial (# of Q)	1000	1000	1000	1000	1000	1000	1667	1667	1667	1667	1667	1667
Generate MA	1000	1000	1000	1000	1000	1000	1435	1392	1352	1532	1539	1482
Generate CMA	1000	1000	1000	1000	1000	1000	1152	1101	1140	1189	1252	1232
CMA filter	922	932	942	944	946	934	1060	1027	1123	1110	1188	1228
Direct Evidence	918	922	938	933	933	931	1042	1009	1002	1079	1171	1173
2 sentence indirect evidence	903	910	922	921	923	914	990	985	965	1038	1129	1116
3 sentence indirect evidence	907	897	925	918	924	920	991	982	980	1041	1125	1122
intersection of 2&3 sentence evidence	901	895	917	911	918	913	976	972	941	1025	1108	1059

Table 7: The dataset scale at each step across popQA and NQ dataset. “intersection of 2&3 sentence evidence” is the count for indirect evidence.

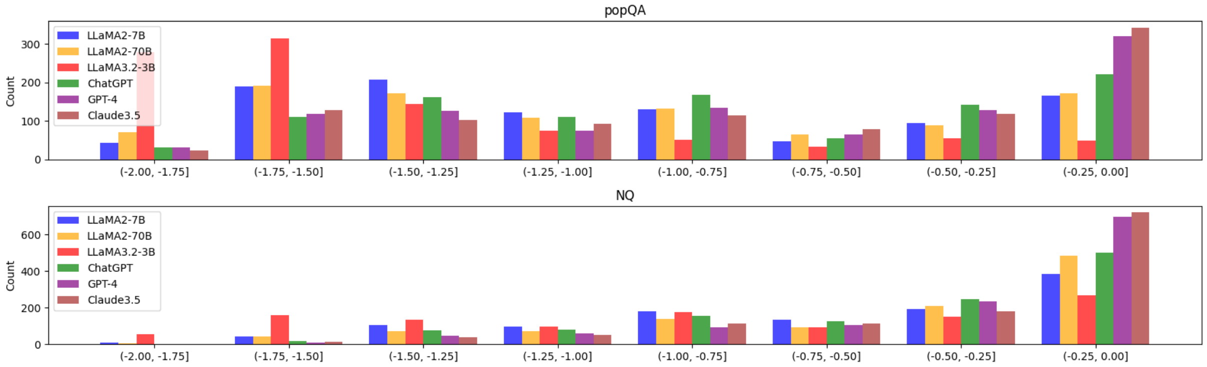


Figure 6: Memory Strength Distribution Across popQA and NQ Datasets. Each dataset is classified into 8 bins. The x-axis shows the range of memory strength for each bin. The y-axis shows the question count in each bin. The NQ dataset exhibits higher overall memory strength. Additionally, larger models (e.g., GPT-4) show stronger memory strength compared to smaller models.

under two evidence styles: Direct Evidence with one sentence and Direct + Paraphrase with three sentences. The results are presented in Table 8. The results show that, for different evidence styles, R_m is higher in the “CMA first” compared to the “MA first”. Comparing the results under the “CMA first”, the R_m of Direct + Paraphrase with three sentences is significantly lower than that of Direct Evidence with one sentence. This demonstrates that paraphrasing direct evidence is an effective method for decreasing R_m. Our conclusion remains unchanged.

B.3 Entity Type

To investigate how entity type influences context faithfulness, we categorize NQ questions into three types based on their expected answer entities: PER (person), LOC (location), and TIM (time), following the classification summarized in Table 3. Specifically:

- **PER** includes questions such as “who”, “what_name”, “who_writes”, “who_sings”, “who_wins”, and “who_plays”.
- **LOC** includes “where” and “which_country”.
- **TIM** includes “when”, “what_year”, and “which_year”.

We then analyze the MA ratio (R_m) across different entity types using ChatGPT and LLaMA2-7B as representative models. For each category, we compute the question count, average memory strength, and MA ratio. Results are shown in Table 9. We observe the following patterns:

- For PER and LOC, which involve textual entities, higher memory strength corresponds to higher MA ratios. This is consistent with the trend reported in Section 4.2.1, where stronger memory increases the likelihood of memory-aligned responses.
- For TIM, which involves numerical entities, the MA ratio is substantially lower than for textual entities. This finding aligns with observations made in Longpre et al. (2021).
- We further divided TIM into four groups based on memory strength levels: low, mid-low, mid-high, and high. The results shown in Table 10 and these results demonstrate a positive correlation between memory strength and MA ratio, which aligns with the conclusion presented in Section 4.2.1.

These results suggest that entity type, particularly the distinction between textual and numerical

Dataset	Evidence Style	S #	LLaMA2-7B			LLaMA2-70B			LLaMA3.2-3B			ChatGPT			GPT-4			Claude3.5		
			$R_m \downarrow$	$R_c \uparrow$	R_u	$R_m \downarrow$	$R_c \uparrow$	R_u	$R_m \downarrow$	$R_c \uparrow$	R_u	$R_m \downarrow$	$R_c \uparrow$	R_u	$R_m \downarrow$	$R_c \uparrow$	R_u	$R_m \downarrow$	$R_c \uparrow$	R_u
popQA	MA first																			
	Direct	1	0.44	99.56	0.0	1.08	98.7	0.22	7.69	85.15	7.16	3.32	94.75	1.93	13.29	84.57	2.14	0.65	83.31	16.04
	Direct + Paraphrase	3	0.11	99.89	0.0	0.43	99.35	0.22	1.07	97.86	1.07	1.39	98.28	0.32	1.29	98.5	0.21	0.11	98.06	1.83
	CMA first																			
	Direct	1	59.37	40.41	0.22	5.75	93.27	0.98	18.8	69.12	12.07	6.22	91.96	1.82	21.44	75.24	3.32	0.43	76.96	22.61
	Direct + Paraphrase	3	17.49	82.51	0.0	1.74	98.05	0.22	14.64	79.92	5.45	1.82	97.75	0.43	2.79	96.78	0.43	0.21	96.88	2.91
NQ	MA first																			
	Direct	1	7.2	92.8	0.0	3.07	96.93	0.0	26.41	59.88	13.71	19.46	75.16	5.38	50.04	47.99	1.96	1.96	56.4	41.64
	Direct + Paraphrase	3	3.26	96.74	0.0	1.19	98.61	0.2	9.38	83.27	7.36	9.55	86.75	3.71	11.27	87.28	1.45	0.26	93.09	6.65
	CMA first																			
	Direct	1	22.26	77.73	0.0	19.13	80.38	0.5	41.33	42.44	16.23	34.48	61.82	3.71	49.19	47.99	2.82	4.78	39.85	55.38
	Direct + Parapharse	3	4.8	95.11	0.1	8.72	90.39	0.89	24.19	65.62	10.18	18.63	78.96	2.41	17.76	80.02	2.22	1.28	78.84	19.88

Table 8: Results of LLM Receptiveness to Different Evidence Styles Across popQA and NQ Datasets. The table presents the MA ratio (R_m), CMA ratio (R_c), and UCT ratio (R_u) for Direct Evidence and Direct + Paraphrase Evidence with CMA first and MA first scenarios. All the ratios are in %.

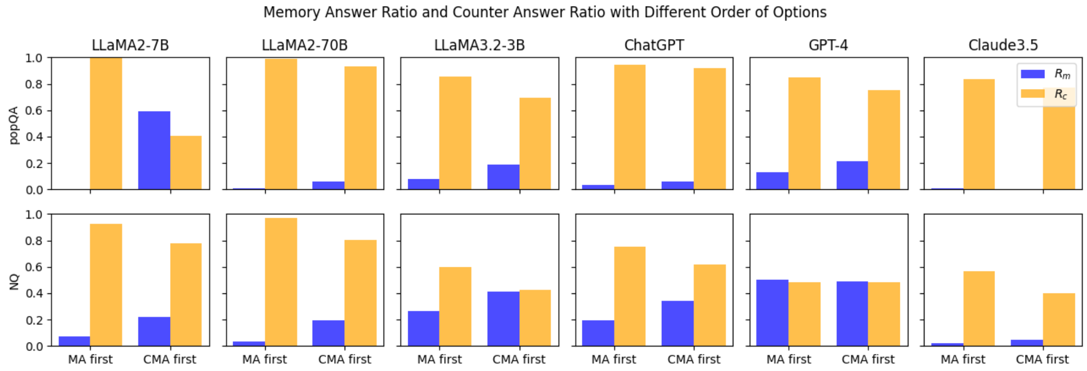


Figure 7: Impact of Option orders on Memory and Counter Ratios Across NQ and popQA Datasets. Either the memory answer ("mem first") or the counter answer ("ctr first") is introduced first to six models.

LLM	Entity	Count	Mem. Strength	MA Ratio (%)
ChatGPT	PER	664	-0.4467	24.54
	LOC	107	-0.5033	19.62
	TIM	217	-0.3244	9.21
LLaMA2-7B	PER	680	-0.6149	8.52
	LOC	95	-0.4344	9.47
	TIM	174	-0.4359	1.15

Table 9: Impact of different entity types (PER: Person, LOC: Location, TIM: Time) on memory strength and MA ratio for ChatGPT and LLaMA2-7B.

LLM	low	mid-low	mid-high	high
ChatGPT	4.76	9.09	9.09	10.08
LLaMA2-7B	0.00	2.13	0.00	2.53

Table 10: MA ratios (%) on TIM-type questions across different memory strength levels.

entities, affects context faithfulness, and that memory strength remains a key influencing factor across different entity types.

B.4 Case Study

Poor reasoning ability may be one of the reasons that LLMs provide MA. In our case study, we prompt LLMs to provide a rationale alongside the answer. We discover that some LLMs seem to exhibit reasoning errors. Below is an example that shows this phenomenon.

Information: Michigan State University holds the record for the highest number of Big Ten football championships, with a total of 36 titles.

Question: who has the most big ten championships in football

Options:

A: Ohio State University has the most Big Ten championships in football with 36 titles.

B: Michigan State University has the most Big Ten championships in football with 36 titles.

C: Uncertain.

Answer:

A: Ohio State University has the most Big Ten championships in football with 36 titles.

According to the information provided, Michigan State University holds the record for the highest number of Big Ten football championships, with a total of 36 titles. Therefore, option A is the correct answer.

In this case, option A represents MA, and option B represents CMA. While the model select MA as the final answer, its rationale indicates that it successfully received and processed the information from CMA. However, for some unknown reason, it still provides the MA as the final answer. This reasoning process is highly confusing. We suspect that the cause of MA can be attributed to LLMs' weakness in reasoning ability. This finding raises an interesting question about the relationship between reasoning ability and memory strength of LLMs. We leave this for future work.

B.5 Impact of Memory Strength with Different Evidence Styles

To demonstrate the relationship between context-faithfulness and memory strength with other evidence styles, we categorize the questions in each dataset into four groups according to the memory strength intervals [-2, -1], (-1, -0.5], (-0.5, -0.25] and [-0.25, 0], The evidence styles are direct + paraphrase evidence with two sentences and indirect evidence with two sentences. Figures 8,9 show the result. The figures show that there is a clear positive correlation between memory strength and MA ratio for both evidence styles, which implies that this positive correlation between memory strength and MA ratio is general.

To demonstrate the relationship between context-faithfulness and memory strength with "CMA first" scenario, we show MA, CMA, and UCT ratios with direct evidence with one sentence under "CMA first" scenario in Figure 10. The positive correlation between memory strength and MA ratio stays unchanged.

C Prompts

In Table 11, we present a detailed list of all the prompts used throughout this study.

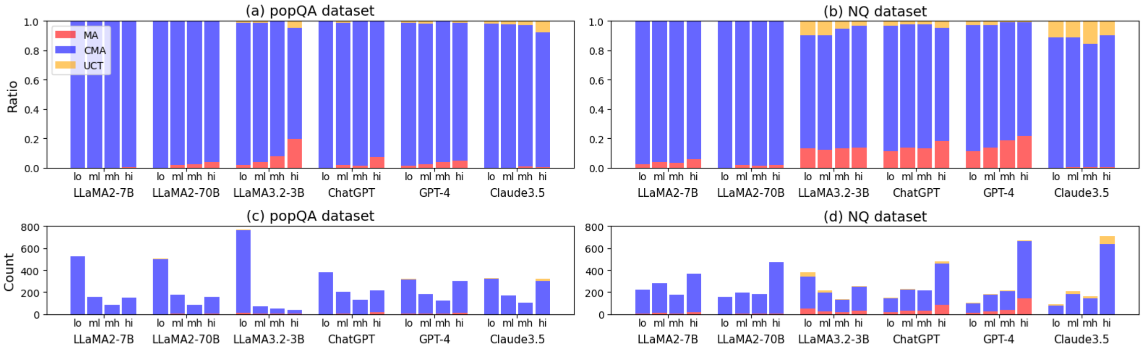


Figure 8: Relationship between Memory Strength and Different Answers' Ratios with Direct + Paraphrase Evidence with Two Sentences.

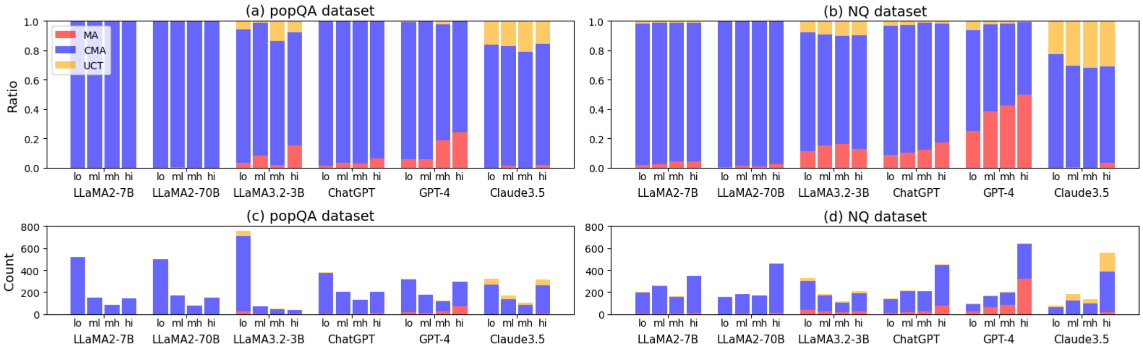


Figure 9: Relationship between Memory Strength and Different Answers' Ratios with Indirect Evidence with Two Sentences.

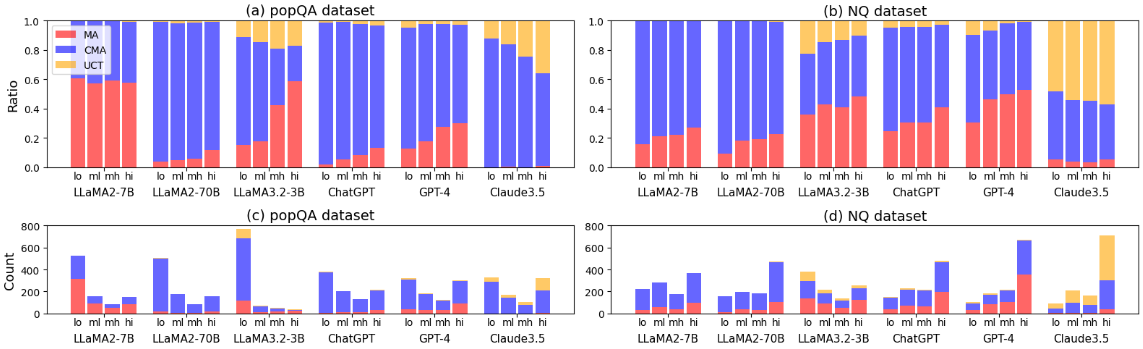


Figure 10: Relationship between Memory Strength and Different Answers' Ratios with Direct Evidence. The option order is Counter First.

Step	index	Prompt Name	Prompts
Step 1: Memory Strength	1	Question paraphrase prompt	Generate 7 meaningful paraphrases for the following question: [Question]. Read the question carefully. Paraphrases:
	2	Question equivalent check prompt	Determine whether the paraphrased question describes the same thing as the original question, and give "Contradicted" if they are not the same, otherwise give "Same" as the result. Q1: [Paraphrased Q1] Q2: [Paraphrased Q2] Keep the answer short and concise.
	3	Answer consistency check prompt	Determine whether the answer ‘A1’ is ‘Contradicted’ or ‘Same’ with the answer ‘A2’ for the question ‘Q’. You need to check whether the two answers exactly have the same answer to the question. The answer could be person, name, place, time, number, genre, occupation, sport, entity, digit, or arithmetical results. If the two answers are the same, give “Same”, otherwise give “Contradicted” as the result. Q: [question] A1: [LLM answer 1] A2: [LLM answer 2] Keep the answer short and concise.
Step 2: MA and CMA	4	Close book QA prompt	Answer the question with one sentence with object and subject only. Give a statement that is most likely to be true directly. Question: [Question] Answer:
	5	Change MA to CMA prompt	Context: [CMA] Change the [entity type] part of the context. When multiple parts need to be changed, only choose one part to change. Answer:
Step 3: Evidence	6	Direct evidence prompt	Please paraphrase the following sentence by changing the terms, order, and phrases, but keep the meaning the same. Sentence: [CMA]
	7	Indirect evidence prompt	Given a claim, please write a short piece of detailed evidence to support it. Please ignore the correctness of the claim. You can make up fake content and supporting evidence but it should be as realistic as possible. Claim: [counter memory answer] Evidence: Give the answer in [2 or 3] sentences directly.
Step 4: Evaluation	8	Evaluate with evidence prompt	According to the given information, choose the best choice from the following options. Information: [evidence] Question: [question] Option: A: [option 1] B: [option 2] ... Answer:

Table 11: Prompts for LLMs in this paper. “[PLACEHOLDER]” is the corresponding input.