Unveiling Factual Recall Behaviors of Large Language Models through Knowledge Neurons

Yifei Wang[*1,2] **,Yuheng Chen**[*2] **,Wanting Wen**[1] **,Yu Sheng**[1,2]**,Linjing Li** [†1,2,3]**, Daniel Zeng**[1,2]
[1] State Key Laboratory of Multimodal Artificial Intelligence Systems,
Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3] Beijing Wenge Technology Co., Ltd, Beijing, China
{wangyifei2022, chenyuheng2022}@ia.ac.cn
{wanting.wen, shengyu2021, linjing.li, dajun.zeng}@ia.ac.cn

Abstract

In this paper, we investigate whether Large Language Models (LLMs) actively recall or retrieve their internal repositories of factual knowledge when faced with reasoning tasks. Through an analysis of LLMs' internal factual recall at each reasoning step via Knowledge Neurons, we reveal that LLMs fail to harness the critical factual associations under certain circumstances. Instead, they tend to opt for alternative, shortcut-like pathways to answer reasoning questions. By manually manipulating the recall process of parametric knowledge in LLMs, we demonstrate that enhancing this recall process directly improves reasoning performance whereas suppressing it leads to notable degradation. Furthermore, we assess the effect of Chain-of-Thought (CoT) prompting, a powerful technique for addressing complex reasoning tasks. Our findings indicate that CoT can intensify the recall of factual knowledge by encouraging LLMs to engage in orderly and reliable reasoning. Furthermore, we explored how contextual conflicts affect the retrieval of facts during the reasoning process to gain a comprehensive understanding of the factual recall behaviors of LLMs.

1 Introduction

Recent advancements in Large Language Models have underscored their exceptional *reasoning* prowess with natural language understanding across a broad spectrum of tasks (Chen et al., 2023a; Kojima et al., 2022; Brown et al., 2020; Creswell et al., 2023). However, amidst these achievements, a specific form of reasoning has been somewhat overlooked and insufficiently investigated: reasoning tasks that require the utilization of internal factual knowledge associations. For instance, when presented with a 2-hop question such as "Who is the chairperson of the manufacturer of the Holden Caprice?" in Figure 1, LLMs must first identify that the manufacturer of the Holden Caprice is General Motors, and subsequently retrieve the name of General Motors' chairperson from their internal knowledge, also referred to as parametric knowledge (Neeman et al., 2023; Zhong et al., 2024). Previous work has shown factual knowledge emerges in both GPT (Meng et al., 2022) and Bert models (Petroni et al., 2019; Jiang et al., 2020). Unlike mathematical (Floyd, 2007) and logical reasoning (Pan et al., 2023), factual reasoning heavily relies on the factual knowledge encoded within LLMs, acquired through extensive pretraining on vast corpora, rather than on user-inputted premises. At the same time, it differs from commonsense reasoning (Zhao et al., 2023; Trinh and Le, 2019), which taps into general knowledge acquired through dynamic training to foster a holistic understanding of the world, instead of emphasizing specific factual information.

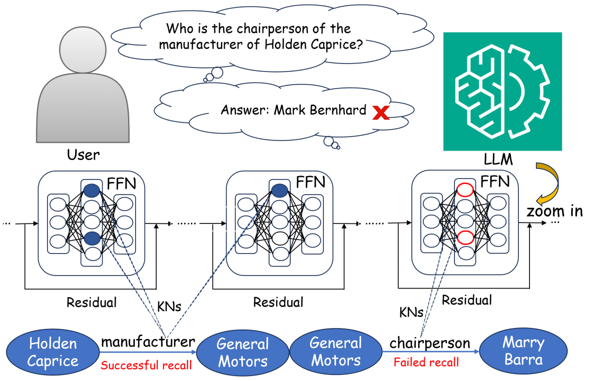


Figure 1: An unsuccessful case of reasoning due to factual retrieval failure of the triplet (General Motors, chairperson, Marry Barra).

Intuitively, it is reasonable to expect LLMs to harness extensive parametric knowledge to tackle reasoning tasks. Yet, an important question emerges: How effectively can LLMs actually retrieve and utilize their internal knowledge for reasoning purposes? Delving into this question is

[*] Equal contribution.
[†] Corresponding author.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 7388–7402
November 12-16, 2024 ©2024 Association for Computational Linguistics

crucial for several reasons. First, efficient use of parametric knowledge may significantly reduce reliance on external data sources, thereby lowering operational costs of data retrieval and API usage. Second, this dynamic capability allows the knowledge within LLMs to flow and interconnect (Onoe et al., 2023), showcasing these models as organic entities rather than static information repositories (Petroni et al., 2019). From a practical perspective, the accurate retrieval and application of parametric knowledge lead to more reliable and interpretable reasoning, enhancing their utility and trustworthiness in real-world applications.

Transformer-based language models have accumulated substantial knowledge through extensive pretraining (Vaswani et al., 2017). A significant body of recent research has focused on the factuality issues of LLMs (Wang et al., 2023). One stream of this research has concentrated on pinpointing the locations within these models' architectures where factual knowledge is stored and encoded (Meng et al., 2022; Dai et al., 2022; Wallat et al., 2020; Geva et al., 2022, 2021). Simultaneously, there has been a concerted effort to understand the mechanism by which this knowledge is *accessed* during the inference phase (Geva et al., 2023; Yang et al., 2024). Another line of work discusses the balance of the retrieved knowledge and its parametric counterparts (Kwiatkowski et al., 2019; Kandpal et al., 2022; Yu et al., 2023). However, the majority of these studies have either been confined to elementary retrieval tasks, such as recalling a single fact object o from a given triplet (s, r, o), or have not delved into the intricacies of factual knowledge recall and utilization in more advanced challenges, particularly within complex reasoning scenarios. Our work addresses these limitations by examining the inner dynamics of factual recall within LLMs during the two-hop factual reasoning process, providing fresh insights into the behavior of factual recall in reasoning and highlighting avenues for enhancing the robustness and reliability of reasoning through more sophisticated knowledge utilization strategies.

In this work, we investigate the harness of internal knowledge for reasoning through the lens of Knowledge Neurons (KNs). We focus on the basic setting of factual reasoning involving the composition of two facts (for example, "Who is the chairperson of the manufacturer of Holden Caprice?" in Figure 1). To achieve this, we carefully craft two-hop reasoning questions dataset that seamlessly integrates with the KN technique. We assess the level of factual recall at each reasoning step by introducing a novel metric, KN Scores. We examine KN Scores under three conditions of two-hop reasoning: no CoT, zero-shot CoT, and few-shot CoT, unveiling the pitfalls existing in the reasoning process and the enhancement effect of CoT (Wei et al., 2022). Then we conduct targeted interventions on KNs to enhance or suppress the factual retrieval process, finding the contributing impact on reasoning performance. Furthermore, we provide a detailed analysis of factual shortcuts (Ju et al., 2024; Du et al., 2023; Li et al., 2024), potentially caused by redundant information stored in models' parameters within LLMs used for reasoning. Finally, we explore how the presence of knowledge conflict outside LLMs influences the factual recall process. Our findings can be summarized as follows:

- LLMs do not consistently retrieve the pertinent factual knowledge essential for reasoning, with more than a third of reasoning errors stemming from deficiencies in the retrieval of factual associations.
- CoT could remarkably enhance the recall of factual knowledge by facilitating engagement in step-by-step reasoning, thereby reducing the likelihood of shortcuts.
- By enhancing and suppressing the recall process, we demonstrate that successful factual retrieval is a pivotal factor in improving reasoning performance.
- The presence of knowledge conflict in context could enhance the retrieval of the corresponding fact in the reasoning process to a degree.

2 Preliminaries

2.1 Problem Formulation

We represent facts, such as "(Holden Caprice, manufacturer, General Motors)", as a triplet (s, r, o), where s is the subject, r is the relation, and o is the object. We formulate two-hop factual reasoning questions as a composition of two linked facts $((s, r_1, o_1), (o_1, r_2, o_2))$, with a bridge entity o_1 connecting them. To query LLMs, these triplets must be converted into natural language queries. For a single relation r, we instruct ChatGPT (gpt-3.5-turbo) to generate query templates as $QT_r(\cdot)$. For instance, the single-relation triplet (Holden Caprice, manufacturer, General Motors) can be converted as $QT_{manufacturer}(HoldenCaprice)$: "Which company manufactures Holden Caprice?".

Similarly, for a composition of two relations r_1 and r_2, we prompt ChatGPT to generate a query template as $QT_{r_2}(r_1(\cdot))$, with $r_1(\cdot)$ denoting the description of the entity related to s via r_1 relation (e.g. The manufacturer of Holden Caprice). We refer to the single-hop query as QT_{1H} and the two-hop query as QT_{2H}.

We consider an autoregressive language model $F : X \to Y$, which accepts an input $x \in X$ and produces a prediction $y \in Y$, continuing the input x. We deem that the model "knows" a fact (s, r, o) if the output $F(QT_r(s))$ matches the ground label o and that LLMs can reason a question involving two-hop fact triplets $((s, r_1, o_1), (o_1, r_2, o_2))$ successfully if the output $F(QT_{r_2}(r_1(s))$ matches the ground label o_2. It is noteworthy that query templates, even for the same single relation, are generated with diversity by ChatGPT. This diversity discourages models from making predictions based on the occurrence of specific words, ensuring that they recall knowledge from within themselves instead. We denote the set of two-hop factual questions as Ω, with Ω_T representing the subset of questions that LLMs can answer correctly and Ω_F denoting the subset of questions that LLMs cannot answer correctly. For simplicity, we use ζ to denote $((s, r_1, o_1), (o_1, r_2, o_2))$, thus we have:

$$\Omega_T = \left\{ \zeta \mid \mathrm{F}_\theta(\mathrm{QT}_{r_2}(r_1(s))) = o_2, \forall \zeta \in \Omega \right\} \quad (1)$$

$$\Omega_F = \left\{ \zeta \mid \mathrm{F}_\theta(\mathrm{QT}_{r_2}(r_1(s))) \neq o_2, \ \forall \zeta \in \Omega \right\} \quad (2)$$

2.2 Knowledge Neurons

Pretrained language models store vast amounts of factual knowledge and have a strong ability to recall this factual knowledge without further training (Petroni et al., 2019; Jiang et al., 2020). Drawing inspiration from the key-value-memory nature of feed-forward layers (Geva et al., 2021), Dai et al. (2022) proposes that factual knowledge is stored in specific neurons within the Feed-Forward Networks (FFNs) of the Transformer models, termed as knowledge neurons. They find that knowledge neurons are activated by knowledge-expressing prompts. The higher the activation of these knowledge neurons is, the more significantly their corresponding facts are expressed. Therefore, to assess the recall and utilization of the fact triplet (s, r, o) necessary in the reasoning process, we refer to the activity of KNs as an indicator of factual recall. We make the following **invariant assumptions**: the KNs responsible for the expression of particular relational facts remain consistent across different application contexts. A specific fact is indicated by the same set of KNs under both single-hop queries and reasoning queries, which is a cornerstone for subsequent experiments. In Appendix B, We detail a methodology that utilizes integrated gradient (Sundararajan et al., 2017) method to compute the contribution of all neurons in the intermediate layers of FFNs to the correct prediction of a multi-token ground truth, identifying neurons with greater contributions as KNs.

3 TFRKN: Two-hop Factual Reasoning for Knowledge Neurons

To investigate the behavior of factual recall in reasoning tasks for LLMs, we have developed a specialized dataset for knowledge neurons called **T**wo-hop **F**actual **R**easoning for **K**nowledge **N**eurons, TFRKN.

Dataset Construction Our dataset consists of two-hop factual questions, where each question involves two facts that are connected by an intermediate entity. LLMs are more likely to recall triplets related to popular entities(Mallen et al., 2023). Therefore, for entity selection, we use the cumulative pageview count over the past 12 months as a metric and select the top 500 popular entities from Wikidata (Vrandečić and Krötzsch, 2014) based on this criterion. Two-hop fact triplets are then extracted from sub-graphs consisting solely of a set of manually selected relations and entities. To identify KNs for each-hop fact, we reformulate each fact triplet into more than five varied natural questions using ChatGPT (Appendix A). The TFRKN dataset encompasses 4,550 distinct instances covering 213 unique relational combinations with a sample instance shown in Table 6.

4 Diagnose the Pitfalls of Factual Recall in Reasoning

In the realm of two-hop factual reasoning, an optimal and dependable reasoning trajectory is a multi-hop reasoning approach (Welbl et al., 2017; Ju et al., 2024). This process requires identifying the bridge entity first and then using it to solve the second hop question, necessitating that LLMs recall the relevant fact at each hop step by step, culminating in the formulation of the correct answers. In this section, we investigate whether LLMs faithfully retrieve factual knowledge at each hop when undertaking reasoning tasks.

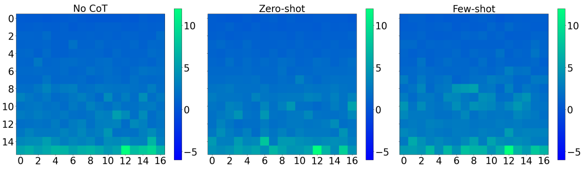


Figure 2: Scaled visualization of neuron activities within the intermediate layers of FFNs in Mistral-7B for the same case (A 32-layer×14336-neuron matrix). The vertical axis shows the depth of layers, while the horizontal axis shows the neuron index in the FFN's intermediate layers. It is evident that KNs are distributed in the middle and final layers.

Models	**Mistral-7B**		**LLaMA2-7B**		**LLaMA3-8B**	
	$\overline{\omega}_1$	$\overline{\omega}_2$	$\overline{\omega}_1$	$\overline{\omega}_2$	ω_1	$\overline{\omega}_2$
Single-hop	2.44	2.61	2.01	1.89	1.70	1.72
	$\Delta_{\overline{\omega}_1}$	$\Delta_{\overline{\omega}_2}$	$\Delta_{\overline{\omega}_1}$	$\Delta_{\overline{\omega}_2}$	$\Delta_{\overline{\omega}_1}$	$\Delta_{\overline{\omega}_2}$
No CoT	-10.84	-11.77	-13.18	-8.18	-10.79	-8.96
Zero-shot	11.56	-8.48	-2.49	-8.30	11.19	6.24
Few-shot	17.36	2.42	1.32	2.46	13.00	7.31

Table 1: KN Scores for three conditions across three models. $\overline{\omega}$ is the KN Score of a specific fact while Δ indicates the change ratio (in percentages) of values compared with the single-hop baselines.

4.1 KN Scores

To evaluate the efficacy of factual recall within LLMs during reasoning tasks, we devise a novel metric termed KN Scores as follows:

$$\mathrm{FFN}^{(l)}(\mathrm{H}^{(l)}) = \mathrm{W}_2^{(l)}\,\mathrm{SiLU}(\mathrm{H}^{(l)}\,\mathrm{W}_1^{(l)}) \quad (3)$$

$$\omega_i^l = \mathrm{SiLU}(\mathrm{H}^{(l)}\,\mathrm{W}_1^{(l)})[i], \quad \forall \omega_i^l \in \omega \quad (4)$$

$$\text{KN Scores} = \frac{1}{|\omega|}\sum \omega_i^l, \forall \omega_i^l \in \omega \quad (5)$$

where $\mathrm{H}^{(l)}$ represents the input to the FFN of the l-th layer, which consists of the outputs from the l-th attention layer combined with the residual stream; ω_i^l denotes the i-th neuron in the l-th intermediate layer of FFN; ω represents the KNs associated with a specific fact triplet, denoted as (s, r, o); $|\omega|$ denotes the size of the set, i.e., the number of KNs; and SiLU denotes the activation function. For the first-hop and second-hop fact, we designate their respective sets of KNs as ω_1 and ω_2. Under the context of a single-hop query, we denote KN Scores as $\{\overline{\omega}|QT_{1H}\}$. Similarly, within the two-hop reasoning context, KN Scores are represented as $\{\overline{\omega}|QT_{2H}\}$.

4.2 Experiment

Setup We begin by filtering out reasoning questions where LLMs are unable to recall all individual facts, ensuring that any reasoning failures are due to the models' inability to retrieve factual information rather than a lack of the foundational knowledge necessary for performing reasoning tasks. We then proceed to employ Fact_1Query and Fact_2Query (in Table 6) from each data point to pinpoint the positions of KNs for each-hop fact. Then we hook the values of each neuron belonging to ω_1 and ω_2 across various query scenarios to compute KN Scores. Using the KN Scores metric, we evaluate the recall of each fact under three distinct experimental conditions: **no CoT**, **zero-shot CoT**, and **few-shot CoT**. For each condition, we record KN Scores for both the first-hop $\{\overline{\omega}_1|QT_{2H}\}$ and the second-hop $\{\overline{\omega}_2|QT_{2H}\}$ facts within the context of two-hop reasoning questions. We select the KN Scores $\{\overline{\omega}_1|QT_{1H}\}$ and $\{\overline{\omega}_2|QT_{1H}\}$ under single-hop queries as baselines since KNs are significantly active in that straightforward context. We experiment with the instructed versions of three popular open-source models: LLaMA2-7B (Touvron et al., 2023), LLaMA3-8B, Mistral-7B (Jiang et al., 2023) (see Appendix C for more experimental details).

4.3 Results

Single-hop vs. Muti-hop Reasoning In reasoning scenarios, LLMs access their internal knowledge less frequently in comparison to the straightforward retrieval of single-hop facts. Table 1 illustrates a notable decrease in KN Scores for all single-hop facts when addressing two-hop reasoning questions. This observation strongly indicates that, in reasoning contexts, LLMs tend to either fail to recall the bridge entity or struggle to identify the second-hop relation, leading to the failure of executing the remaining multi-hop reasoning as anticipated. Compared to directly recalling single-hop facts (e.g., "Who is the chairperson of General Motors?"), it is more challenging for LLMs to recall and organize relevant facts for reasoning. LLMs may take alternative salient pathways existing in their parameters, such as shortcuts, rather than engaging in systematic, step-by-step reasoning.

CoT vs. No CoT CoT, whether zero-shot or few-shot, markedly improves factual knowledge utilization in LLMs over no CoT (KNs are more activated under CoT settings in Figure 2), which is evidenced by a higher $\Delta_{\overline{\omega}_1}$ and $\Delta_{\overline{\omega}_2}$ compared with no CoT

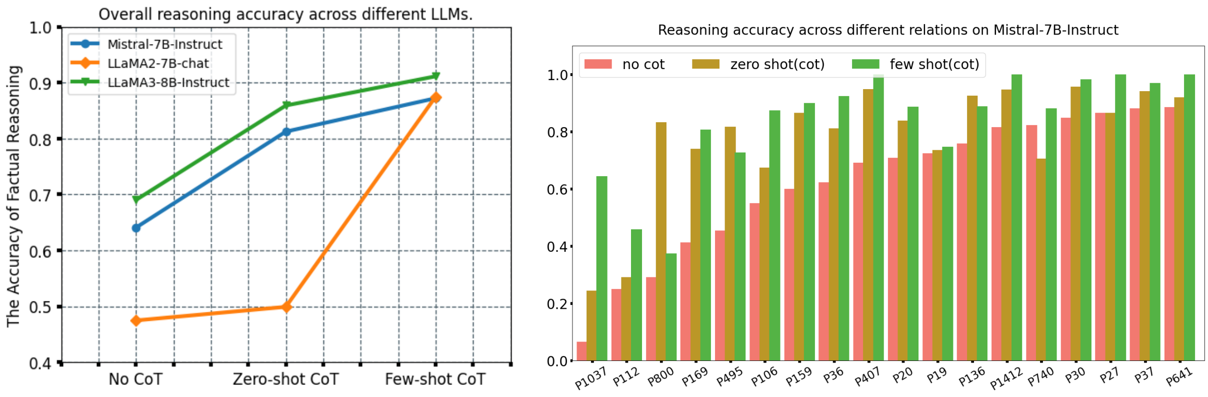


Figure 3: Overall reasoning performance on TFRKN under different CoT situations.

setting, as shown in Table 1. We posit that **this enhancement is likely driven by the step-by-step thinking process, which further stimulates the recall of facts as multi-hop reasoning progresses.** This hypothesis can be supported by comparing the zero-shot and few-shot CoT settings. Across three models, it is clear that zero-shot CoT struggles to significantly improve the recall of the second-hop fact compared to the reinforcement of the first-hop fact recall. However, consistent improvement across both triplets can be observed for few-shot settings. This observation strongly suggests that the reasoning direction in zero-shot scenarios is unclear, which prevents models from effectively identifying which relations of facts concerning the bridge entity to retrieve. In stark contrast, few-shot scenarios often mitigate this issue. Through the acquisition of knowledge from contextual demonstrations, models are more inclined to determine the subsequent phase in the reasoning trajectory and, in turn, adeptly utilize the relevant factual information via their attention mechanisms.

Factual Recall vs. Reasoning Accuracy The combination of Figure 3 and Table 1 illustrates a positive correlation between the recall of relevant fact triplets and reasoning accuracy. This relationship is especially pronounced in the case of LLaMA3-8B model under few-shot CoT, where the maximum increase in the recall of both $\Delta_{\overline{\omega}_1}$ and $\Delta_{\overline{\omega}_2}$ leads to the highest reasoning accuracy. However, the eliciting effect of CoT on factual recall across various LLMs is not uniform. For instance, zero-shot CoT mitigates the forgetting of factual information to some extent for LLaMA2-7B, whereas for LLaMA3-8B, zero-shot CoT enhances the retrieval of factual information to a level comparable to few-shot CoT. This adequately illustrates that the efficacy of CoT is also contingent upon the intrinsic capabilities of the LLMs themselves.

5 Interventions on the Recall of Facts

5.1 Enhance and Suppress KNs

To gain a deeper understanding of factual recall behaviors, we intervene in the retrieval of specific knowledge within LLMs by manually adjusting the activation levels of KNs. Specifically for each factual triplet (s, r, o), we modulate the internal recall by adjusting the values of the KNs associated with this triplet, either amplifying or diminishing them according to Equation 6.

$$\begin{cases} \text{Enhance:}\, \omega_i^l = n \times \omega_i^l, n > 1, \forall \omega_i^l \in \omega \\ \text{Suppress:}\, \omega_i^l = 0, \quad \omega_i^l \in \omega \end{cases} \tag{6}$$

5.2 Experiment

Setup We have meticulously designed four sets of controlled experiments on TFRKN to monitor changes in reasoning outcomes. The experimental paradigms are as follows: (1) Base: We allow LLMs to respond to two-hop questions under standard conditions (2) Enhance: For questions answered incorrectly under Base situation, we amplify the activation level of KNs and subsequently assess the reasoning accuracy. (3) Suppress: Conversely, for two-hop questions correctly answered in the Base scenario, we reduce the activation of relevant KNs and evaluate the reasoning accuracy afterward. (4) Random: To establish a baseline for comparison with conditions (2) and (3), we randomly select an equal number of neurons and enhance or suppress their activation accordingly, facilitating a comparative analysis.

	Mistral-7B		LLaMA2-7B		LLaMA3-8B	
Base	64.09	–	47.48	–	69.03	–
Enha.	Δ	**ER**	Δ	**ER**	Δ	**ER**
ω_1	3.92	18.19	8.79	19.58	4.48	21.24
ω_2	6.16	28.57	13.15	30.39	7.28	34.51
ω_{12}	15.11	**31.05**	15.30	**34.97**	8.02	**38.05**
ω_r	4.57	2.74	7.65	17.79	0.19	0.88
Supp.	Δ	**SR**	Δ	**SR**	Δ	**SR**
ω_1	-20.06	32.28	-18.00	38.07	-24.53	38.07
ω_2	-29.01	46.70	-24.35	50.78	-39.18	53.03
ω_{12}	-49.53	**77.29**	-30.32	**63.85**	-62.59	**91.61**
ω_r	-5.78	9.02	-12.12	25.54	-2.52	3.65

Table 2: Results of the controlled experiments after interventions on ω_1, ω_2 and ω_{12} under no CoT setting. Δ denotes variation in accuracy and ω_r is established as the baseline for enhancing or suppressing KNs of both facts, with ER/SR values expressed as percentages.

Metrics We design a novel metric, termed **Enhance Ratio (ER)**, which serves to quantify the impact of factual retrieval failures on reasoning outcomes. ER is calculated by calculating the percentage of reasoning instances that are initially incorrect but are successfully resolved following the enhancement of KNs as Equation 7. Analogously, we define another metric **Suppress Ratio (SR)** to measure the obstructive effect of suppressed KNs on the reasoning process. The SR is ascertained by evaluating the ratio of cases where correct reasoning is converted to incorrect after the suppression of KNs, as outlined in Equation 8:

$$\mathrm{ER} = \frac{|\{\zeta \mid \mathrm{F}_{\theta'}(\mathrm{QT}_{r_2}(r_1(s))) = o_2\}|}{|\Omega_F|}, \forall \zeta \in \Omega_F \quad (7)$$

$$\mathrm{SR} = \frac{|\{\zeta \mid \mathrm{F}_{\theta''}(\mathrm{QT}_{r_2}(r_1(s))) \neq o_2\}|}{|\Omega_T|}, \forall \zeta \in \Omega_T \quad (8)$$

where θ' denotes the parameters of the enhanced model while θ'' represents the parameters of the suppressed model. $QT_{r_2}(r_1(s))$ represents the reasoning question derived from two-hop fact triplets $((s, r_1, o_1), (o_1, r_2, o_2))$ with the ground truth o_2.

5.3 Results

Finding 1 In Table 2, more than one-third of reasoning failures are caused by issues of factual retrieval. The ER values show a consistent and progressive increase as the interventions progress from targeting ω_1, to KNs associated with the second-hop ω_2, and ultimately to a combined intervention on both, ω_{12}. This pattern indicates that many initially incorrect answers stem from retrieval failure of either the first hop, the second hop, or both during the reasoning process. Additionally, recalling the second-hop facts is more challenging for LLMs, as shown by the higher ER after enhancing ω_2 compared to ω_1. Suppressing factual information significantly harms reasoning performance, with accuracy dropping by over 77% on average when both factual elements are suppressed. Therefore, the successful retrieval of factual associations at each reasoning step is crucial for correct reasoning.

Models	**Mistral-7B**		**LLaMA2-7B**		**LLaMA3-8B**	
Enha.	ω_{base}	ω_{12}	ω_{base}	ω_{12}	ω_{base}	ω_{12}
No CoT	2.74	31.05	17.79	34.97	0.88	38.05
Zero-shot	7.44	53.50	23.36	56.23	23.97	54.79
Few-shot	2.92	39.60	12.51	48.09	2.04	51.02
Supp.	ω_{base}	ω_{12}	ω_{base}	ω_{12}	ω_{base}	ω_{12}
No CoT	9.02	77.29	25.54	63.85	3.65	91.61
Zero-shot	9.76	68.43	10.80	71.80	8.25	74.16
Few-shot	0.11	50.48	5.33	32.09	0.21	65.92

Table 3: ER/SR Results of enhancing and suppressing the expression of both triplets under both CoT and no CoT conditions. In the enhancement scenario, the numbers represent ER metrics, whereas in the suppression scenario, they denote SR metrics.

Finding 2 CoT strengthens a passive internal retrieval of relevant facts, implicitly prompting the expression of factual triplets. Evidence 1: In Table 3, across the scenarios of no CoT, zero-shot CoT, and few-shot CoT, suppression of factual KNs results in $SR_{No\_cot} > SR_{Zero\_shot}$ and $SR_{No\_cot} > SR_{Few\_shot}$, which indicates that CoT likely stimulates the hydra effect (McGrath et al., 2023), which implements actively self-repairing computations to compensate the suppression effects caused by low activation levels of KNs. Evidence 2: Similarly, enhancement of factual KNs results in $ER_{No\_cot} < ER_{Zero\_shot}$ and $ER_{No\_cot} < ER_{Few\_shot}$, which suggests that CoT further stimulates the internal recall process within LLMs, thus strengthening the enhancement effects of KNs. Therefore, CoT indeed can contribute to the recalling process.

6 Analysis of Shortcuts

In this section, we investigate whether successful two-hop reasoning implies the successful recall of factual knowledge. In other words, we examine whether accurate reasoning outcomes stem from a thorough grounding in multi-hop knowledge reasoning or are facilitated by alternative shortcuts.

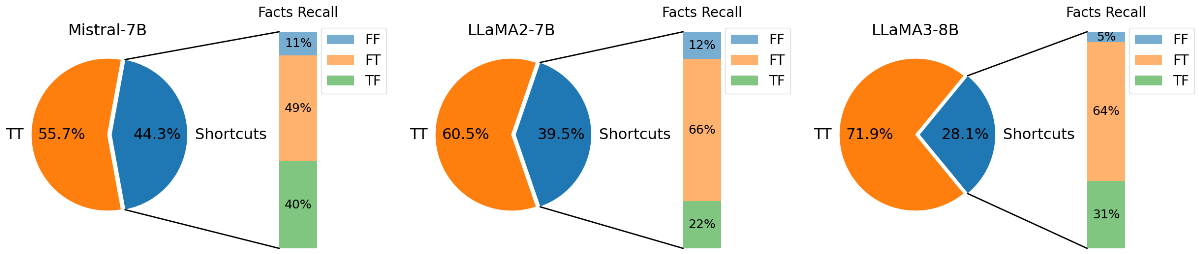


Figure 4: An in-depth analysis of shortcut scenarios under no CoT. TT represents successful recall of both facts.

6.1 Experiment

Setup We investigate the utilization of individual fact triplets in correctly answered two-hop questions by analyzing the KN Scores for each triplet. We compare these scores with those observed during single-hop queries to establish a threshold, denoted as τ, which serves as a benchmark for identifying the effective use of facts in the reasoning process. If the activation level of KNs falls significantly below this threshold in comparison to single-hop queries, this indicates an under-utilization of the corresponding fact. Conversely, if it exceeds the threshold, the fact is considered adequately utilized. Using this criterion, we classified the correctly answered questions into four distinct categories: (1) **FT**: Unsuccessful recall of the first-hop fact but successful second-hop recall; (2) **TF**: Successful first-hop recall but unsuccessful second-hop recall; (3) **FF**: Neither fact successfully recalled and (4) **TT**: Both facts successfully recalled. Except for TT, the other three situations are defined as *Shortcuts*.

Models	Mistral-7B		LLaMA2-7B		LLaMA3-8B	
	MH	SC	MH	SC	MH	SC
No CoT	55.75	44.25	60.51	39.49	71.89	28.11
Zero-shot	70.84	29.16	64.26	35.74	95.66	4.34
Few-shot	**91.23**	**8.77**	**89.02**	**10.98**	**97.65**	**2.35**

Table 4: The fraction of MH and SC in correctly answered examples (TT+FT). MH denotes successful retrieval of both facts while others denote by SC.

6.2 Results Analysis

According to Table 4, under normal conditions, a considerable proportion of correctly answered questions under no CoT setting rely on shortcuts, possibly due to word associations intrinsic to LLMs,as observed by Yang et al. (2024). Notably, the Mistral-7B model stands out for its unexpected reliance on shortcuts to solve over 44 percent of the questions successfully. Even with large-scale models possessing 7 billion parameters, LLMs still rely on certain segments of the reasoning chain to arrive at answers. The introduction of CoT effectively decreases the trend of taking shortcuts by forcing LLMs to recall more relevant facts and engage in multi-hop reasoning. Under few-shot CoT setting, all LLMs solve over 90 percent of questions on average through multi-hop reasoning, reducing the ratio of shortcuts to nearly zero.

Figure 4 provides a closer look at the shortcut phenomenon. The percentage of FF is significantly low, illustrating that it is hard for LLMs to fail to retrieve any factual information relevant when presented with the clues of overlapping entities or relational vocabulary in queries. For most instances of shortcuts, LLMs prefer to utilize the second-hop fact to directly answer reasoning questions, skipping the intermediate reasoning steps and relying on the object o_2 in the second-hop to cheat (a high ratio for FT). For TF cases, there might exist direct associations between the head entity s and the tail entity o_2 leveraged to derive correct answers. In conclusion, experimental results show that recalling two-hop facts (TT) benefits the model's reasoning performance. Specifically, with the presence of CoT, the proportion of TT significantly increases and the model's reasoning accuracy improves substantially.

7 Impact of Contextual Conflict

The capacity of utilizing internal factual knowledge is contingent not solely upon the intrinsic properties of LLMs, but is also significantly influenced by the context within which they operate. This section elucidates how the presence of knowledge conflicts within a given context can impact the mechanisms of the retrieval process during reasoning.

7.1 Experiment

Setup For each data point, we formulate a single-hop conflict fact by devising a set of potential objects denoted as O_{candi} for its r. From this set,

Relational Facts	Type	Examples
⟨ `Middlemarch, author, George Eliot`⟩ ⟨`George Eliot, place of death, London`⟩	Distraction	Context: Carl Sagan works at Cornell University. Question: Where did the author of Middlemarch pass away? A:
	Conflict 1	Context: The author of Middlemarch is Jean Genet. Question: Where did the author of Middlemarch pass away? A:
	Conflict 2	Context: George Eliot died in the city of Atlanta. Question: Where did the author of Middlemarch pass away? A:

Table 5: Knowledge conflict and knowledge distraction examples

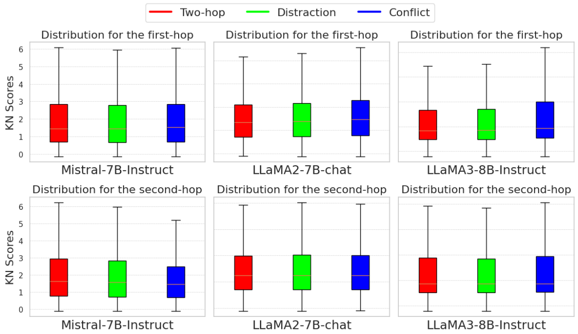


Figure 5: Results of constructing the knowledge distraction and knowledge conflict for the first-hop fact.

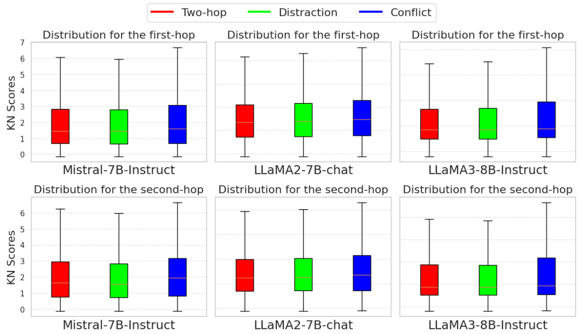


Figure 6: Results of constructing the knowledge distraction and knowledge conflict for the second-hop fact.

we deliberately select an object $o^* \neq o$ to introduce a knowledge conflict. In contradistinction, we also fabricate an entirely unrelated fact for each data point to serve as a distractor, referred to as knowledge distraction (See detailed construction in Appendix D). We then respectively append the knowledge conflict and knowledge distraction sentences before the two-hop question under no CoT setting, which is input into LLMs. Then we observe the values of KN Scores for each-hop fact. The examples of knowledge conflict and distraction for the first-hop and the second-hop facts are shown in Table 5.

7.2 Results Analysis

The presence of knowledge conflict within the context consistently augments the faithfulness of LLMs in the corresponding fact. According to Figure 5 and Figure 6, the context of knowledge conflict results in the highest KN Scores of the corresponding hop fact [1], which indicates counterfactual context significantly improves the internal retrieval of that corresponding hop fact. It illustrates LLMs exhibit greater confidence in their encoded knowledge when confronted with knowledge conflict, a finding that aligns with the studies conducted by Zhou et al. (2023) and Li et al. (2023). When the knowledge presented in the context conflicts with the second-hop fact, it not only reinforces the retrieval of the second-hop fact but also enhances the recall of the first-hop fact. It is plausible that the introduction of the subject o_1 encourages LLMs to recall the precise triplet (s, r_1, o_1). However, this effect does not extend to the first-hop fact. The occurrence of knowledge distraction appears not to cause much obstruction to the factual recall within LLMs. On the contrary, it may even stimulate LLMs to retrieve more facts sometimes, as evidenced by the high KN Scores for the first-hop fact of LLaMA2-7B when the knowledge distractor corresponding to the second-hop fact appears in Figure 6.

8 Related Work

Multi-hop Reasoning Multi-hop reasoning poses a significant challenge for LLMs. Several studies have endeavored to address this challenge through the development of more faithful reasoning techniques (Creswell and Shanahan, 2022; Chen et al., 2023b; Creswell et al., 2023). One such approach is CoT, which stimulates LLMs to produce deductive intermediate steps, fostering a step-by-step analytical process (Chu et al., 2024). Another line of research is focused on visualizing the implicit logical structures within LLMs from the perspective of mechanistic

[1] The results were obtained from a one-tailed paired sample t-test, conducted at a significance level of 0.05.

interpretability (Yang et al., 2024). For example, a recent study by Hou et al. (2023) recovers the reasoning tree from models's attention patterns using MechanisticProbe.

CoT Mechanism A large body of literature is dedicated to the theoretical and empirical exploration of the mechanism underlying CoT (Saparov and He, 2023; Tan, 2023; Feng et al., 2023; Prystawski et al., 2023; Xie et al., 2024). Some research endeavors to delve into a reverse-engineering analysis of CoT prompting, uncovering the intricate information pathways that facilitate the generation of responses (Dutta et al., 2024). However, the majority of these studies concentrate on the rationales produced by CoT and have largely overlooked the broader implications for factual retrieval processes. In our current work, we complement this aspect and present compelling evidence that CoT significantly bolsters the internal recall of factual information.

9 Conclusions

This paper aims to provide a comprehensive understanding of factual recall behaviors for LLMs. We find that a considerable portion of reasoning failures are due to retrieval failures. Manually enhancing the internal recall within LLMs can improve reasoning performance. For LLMs, they not only rely on multi-hop reasoning but also rely on other inference ways in LLMs such as shortcuts. CoT can significantly stimulate LLMs to recall more facts by compelling models to engage in step-by-step thinking, diminishing the possibilities of taking shortcuts. The knowledge conflict existing in context could improve the confidence of parametric knowledge, therefore enhancing the internal recall.

Limitations

While our study provides novel insights into the internal factual recall behaviors of LLMs during reasoning tasks, it is important to acknowledge several limitations.

Generalizability: While the current study is primarily based on specific LLMs and the TFRKN dataset, future research should extend these findings to verify their generalizability across various models and datasets

Theoretical Analysis: Although empirical evidence has been provided through targeted interventions, a deeper theoretical analysis is needed to fully comprehend the underlying reasons for the observed phenomena.

Practical Applications: The paper discusses theoretical aspects and potential improvements in reasoning accuracy but does not delve into how these findings can be applied in practical scenarios to enhance the reasoning capabilities of LLMs.

Impact of Contextual Factors: While the paper touches upon the influence of contextual conflicts on knowledge retrieval, a more comprehensive analysis of various contextual factors and their impact on reasoning is needed.

Acknowledgements

This work was supported in part by the Strategic Priority Research Program of Chinese Academy of Sciences under Grant #XDA27030100 and the National Natural Science Foundation of China under Grants #72293573.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2023a. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*.

Zeming Chen, Gail Weiss, Eric Mitchell, Asli Celikyilmaz, and Antoine Bosselut. 2023b. Reckoning: Reasoning through dynamic knowledge encoding. In *Advances in Neural Information Processing Systems*, volume 36, pages 62579–62600. Curran Associates, Inc.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2024. Navigate through enigmatic labyrinth a survey of chain of thought reasoning: Advances, frontiers and future. *Preprint*, arXiv:2309.15402.

Antonia Creswell and Murray Shanahan. 2022. Faithful reasoning using large language models. *Preprint*, arXiv:2208.14271.

Antonia Creswell, Murray Shanahan, and Irina Higgins. 2023. Selection-inference: Exploiting large language models for interpretable logical reasoning. In *The Eleventh International Conference on Learning Representations*.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.

Mengnan Du, Fengxiang He, Na Zou, Dacheng Tao, and Xia Hu. 2023. Shortcut learning of large language models in natural language understanding. *Commun. ACM*, 67(1):110–120.

Subhabrata Dutta, Joykirat Singh, Soumen Chakrabarti, and Tanmoy Chakraborty. 2024. How to think step-by-step: A mechanistic understanding of chain-of-thought reasoning. *Preprint*, arXiv:2402.18312.

Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. 2023. Towards revealing the mystery behind chain of thought: A theoretical perspective. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Juliet Floyd. 2007. 75Wittgenstein on Philosophy of Logic and Mathematics. In *The Oxford Handbook of Philosophy of Mathematics and Logic*. Oxford University Press.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235, Singapore. Association for Computational Linguistics.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yifan Hou, Jiaoda Li, Yu Fei, Alessandro Stolfo, Wangchunshu Zhou, Guangtao Zeng, Antoine Bosselut, and Mrinmaya Sachan. 2023. Towards a mechanistic interpretation of multi-step reasoning capabilities of language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4902–4919, Singapore. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How Can We Know What Language Models Know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Tianjie Ju, Yijin Chen, Xinwei Yuan, Zhuosheng Zhang, Wei Du, Yubin Zheng, and Gongshen Liu. 2024. Investigating multi-hop factual shortcuts in knowledge editing of large language models. *Preprint*, arXiv:2402.11900.

Nikhil Kandpal, H. Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2022. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

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

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.

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.

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.

Thomas McGrath, Matthew Rahtz, Janos Kramar, Vladimir Mikulik, and Shane Legg. 2023. The hydra effect: Emergent self-repair in language model computations. *Preprint*, arXiv:2307.15771.

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

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.

Yasumasa Onoe, Michael Zhang, Shankar Padmanabhan, Greg Durrett, and Eunsol Choi. 2023. Can LMs learn new entities from descriptions? challenges in propagating injected knowledge. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5469–5485, Toronto, Canada. Association for Computational Linguistics.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Wang. 2023. Logic-LM: Empowering large language models with symbolic solvers for faithful logical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3806–3824, Singapore. Association for Computational Linguistics.

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

Ben Prystawski, Michael Y. Li, and Noah Goodman. 2023. Why think step by step? reasoning emerges from the locality of experience. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Abulhair Saparov and He He. 2023. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *The Eleventh International Conference on Learning Representations*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Juanhe (TJ) Tan. 2023. Causal abstraction for chain-of-thought reasoning in arithmetic word problems. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 155–168, Singapore. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Trieu H. Trinh and Quoc V. Le. 2019. A simple method for commonsense reasoning. *Preprint*, arXiv:1806.02847.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Jonas Wallat, Jaspreet Singh, and Avishek Anand. 2020. BERTnesia: Investigating the capture and forgetting of knowledge in BERT. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 174–183, Online. Association for Computational Linguistics.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, Yidong Wang, Linyi Yang, Jindong Wang, Xing Xie, Zheng Zhang, and Yue Zhang. 2023. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *Preprint*, arXiv:2310.07521.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed 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*, volume 35, pages 24824–24837. Curran Associates, Inc.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2017. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

Zhihui Xie, Jizhou Guo, Tong Yu, and Shuai Li. 2024. Calibrating reasoning in language models with internal consistency. *Preprint*, arXiv:2405.18711.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. 2024. Do large language models latently perform multi-hop reasoning? *Preprint*, arXiv:2402.16837.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023. Generate rather than retrieve: Large language models are strong context generators. In *The Eleventh International Conference on Learning Representations*.

Zirui Zhao, Wee Sun Lee, and David Hsu. 2023. Large language models as commonsense knowledge for large-scale task planning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Ming Zhong, Chenxin An, Weizhu Chen, Jiawei Han, and Pengcheng He. 2024. Seeking neural nuggets: Knowledge transfer in large language models from a parametric perspective. In *The Twelfth International Conference on Learning Representations*.

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.

A Details of Dataset Construction

A.1 Sampling two-hop factual triples

Our dataset is constructed based on Wikidata (Vrandečić and Krötzsch, 2014), a structurally optimized database covering nearly all domains. The dataset is available at `https://github.com/wangyifei0047/TFRKN`.

First we show **manually selected relations** that are used to construct two-hop relations:

- P30, P36, P35, P1037, 1308, P164, P449, P488, P178, P159, P286, P413, P641, P800, P937
- P136, P106, P495, P740, P37, P407, P170, P50,P364,P112, P108, P175, P27, P40, P69, P19

While LLMs have been shown to store a vast amount of factual knowledge, studies indicate that they are more likely to recall triplets related to popular entities (Mallen et al., 2023). Therefore, when constructing the dataset, we employ the cumulative pageviews count over the past 12 months as a measure and select the top 500 popular entities based on this criterion. Two-hop reasoning chains are then extracted from the sub-graphs consisting solely of the aforementioned relations and entities, like *(Holden Caprice, manufacturer, General Motors), (General Motors, chairperson, Mary Barra).*

A.2 Generating Queries using ChatGPT

Having acquired the triplet format of reasoning queries, our current objective is to transform these triplets into natural language expressions in queries. Moreover, for effective integration of the Knowledge Neuron technique, it is essential to rephrase individual triplets into multiple natural language expressions. As knowledge neurons demonstrate indifference towards specific knowledge representations, employing diverse question formats aids in identifying authentic knowledge neurons. Whether in the formulation of reasoning queries or the generation of individual triplet queries, we capitalize few-shot learning capabilities of ChatGPT (gpt-3.5-turbo) to autonomously generate natural language questions. Concretely, we leveraged few-shot capabilities in LLMs to generate multiple queries for individual fact (s, r, o), as well as reasoning questions from two-hop facts $((s_1, r_1, o_1), (o_1, r_2, o_2))$. For the generation of single-fact queries, we provide relation labels and relation definitions as additional information for LLMs to generate accurate subject-relation queries (Figure 8). For the generation of reasoning questions, two-hop relation labels and explanations are also provided besides four in-context demonstrations (Figure 7).

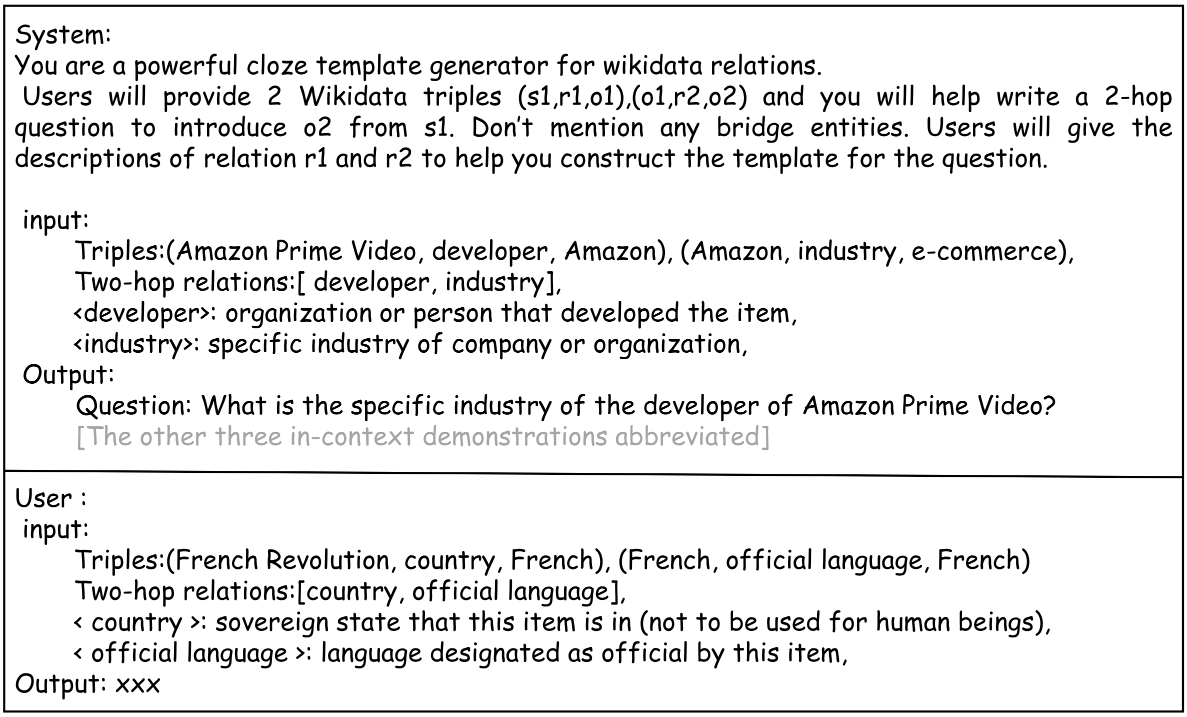
An instance from TFRKN is depicted in Table 6. This approach not only surpasses the limitations imposed by manual templates but also guarantees the production of high-quality and diverse questions. Overall, the dataset comprises 4,550 instances spanning 213 unique combinations of relations.

B Knowledge Neurons

In this part, we detailedly illustrate the methodology of the identification of KNs using the integrated gradient method. Given a specific relational fact: (s, r, o); A set of knowledge-expressing queries (Fact1Query and Fact2Query in Table 6): $< query_1, query_2, \cdots, query_L >$. We define the representation of the i-th neuron in the l-th intermediate layer in FFNs as w_i^l,

$$P_{[t_1,\cdots,t_n],y}(w_i^{(l)}) = P(y|\left[t_1,\cdots,t_n\right], w_i^{(l)} = \tilde{w}_i^{(l)}) \quad (9)$$

where $[t_1, t_2, \cdots, t_n]$ represents the token sequence of inputs, $\tilde{w}_i^{(l)}$ represents the constant value assigned to $w_i^{(l)}$, and Equation 9 denotes the probability of next token y predicted by LLMs, given the token sequence $[t_1, t_2, \cdots, t_n]$ after $w_i^{(l)}$ is assigned the value $\tilde{w}_i^{(l)}$.

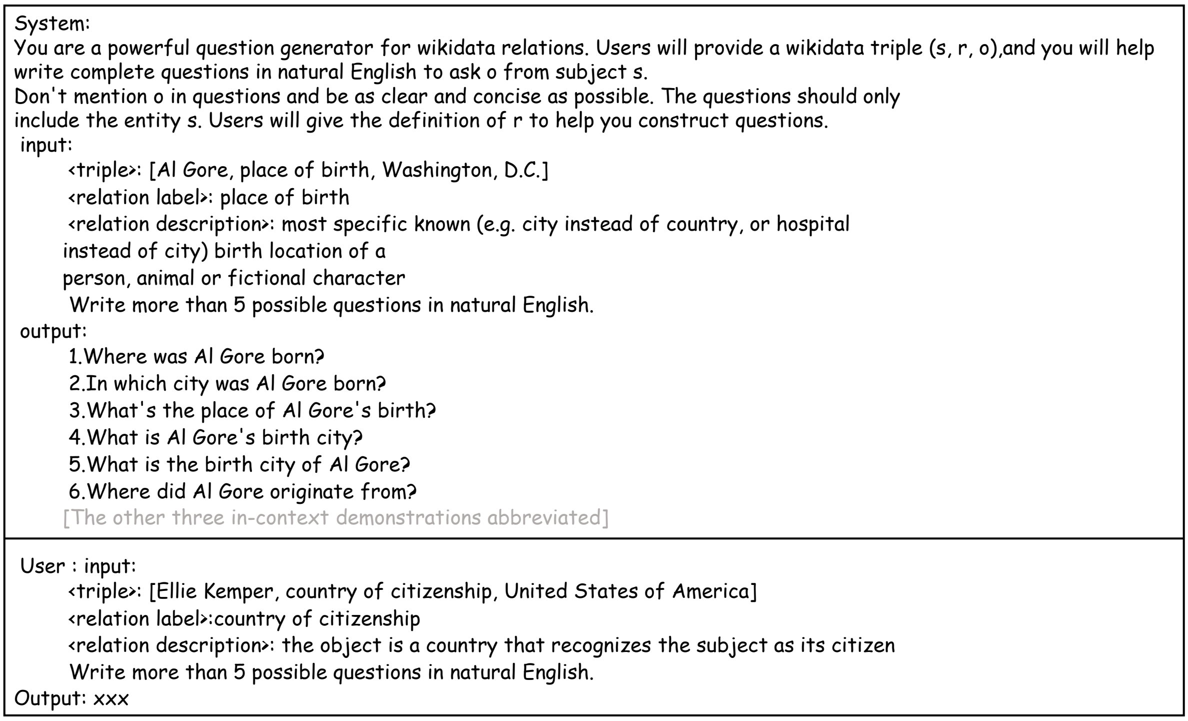


```
System:
You are a powerful cloze template generator for wikidata relations.
 Users will provide 2 Wikidata triples (s1,r1,o1),(o1,r2,o2) and you will help write a 2-hop question to introduce o2 from s1. Don't mention any bridge entities. Users will give the descriptions of relation r1 and r2 to help you construct the template for the question.

 input:
       Triples:(Amazon Prime Video, developer, Amazon), (Amazon, industry, e-commerce),
       Two-hop relations:[ developer, industry],
       <developer>: organization or person that developed the item,
       <industry>: specific industry of company or organization,
 Output:
       Question: What is the specific industry of the developer of Amazon Prime Video?
       [The other three in-context demonstrations abbreviated]

User :
 input:
       Triples:(French Revolution, country, French), (French, official language, French)
       Two-hop relations:[country, official language],
       < country >: sovereign state that this item is in (not to be used for human beings),
       < official language >: language designated as official by this item,
Output: xxx
```

Figure 7: An example of using ChatGPT to generate 2-hop questions from Wikidata triples.



```
System:
You are a powerful question generator for wikidata relations. Users will provide a wikidata triple (s, r, o),and you will help
write complete questions in natural English to ask o from subject s.
Don't mention o in questions and be as clear and concise as possible. The questions should only
include the entity s. Users will give the definition of r to help you construct questions.
 input:
       <triple>: [Al Gore, place of birth, Washington, D.C.]
       <relation label>: place of birth
       <relation description>: most specific known (e.g. city instead of country, or hospital
      instead of city) birth location of a
      person, animal or fictional character
       Write more than 5 possible questions in natural English.
 output:
       1.Where was Al Gore born?
       2.In which city was Al Gore born?
       3.What's the place of Al Gore's birth?
       4.What is Al Gore's birth city?
       5.What is the birth city of Al Gore?
       6.Where did Al Gore originate from?
      [The other three in-context demonstrations abbreviated]

 User : input:
       <triple>: [Ellie Kemper, country of citizenship, United States of America]
       <relation label>:country of citizenship
       <relation description>: the object is a country that recognizes the subject as its citizen
       Write more than 5 possible questions in natural English.
Output: xxx
```

Figure 8: An example of using ChatGPT to generate single-fact queries from triples and relation information(labels and descriptions).

The attribution scores quantify the contribution of individual neurons to correct predictions. By gradually restoring each neuron's value from 0 to its original level, the gradients of the probability of the correct token with respect to each neuron are integrated, as shown in Equation 10.

Equation 10 is applied to the calculation of attribution scores for single-token target o. The method for computing attribution scores for multi-token target o is described in Equation 11. Assuming the tokenized sequence of a relational-fact query and the corresponding ground truth respectively are

Triples	(Holden Caprice, manufacturer, General Motors) (General Motors, chairperson, Mary Barra)
Fact$_1$Query	1. Who or what company manufactures Holden Caprice? 2. What company created Holden Caprice? 3. Who is responsible for making Holden Caprice? 4. What entity produces Holden Caprice? 5. Which organization is behind the production of Holden Caprice?
Fact$_2$Query	1. Who is the chairperson of General Motors? 2. Who is the head of General Motors? 3. Who presides over General Motors as its chairperson? 4. Who currently serves as the chairperson of General Motors? 5. What is the name of the person who chairs General Motors?
Reason_Q	Who is the chairperson of the manufacturer of Holden Caprice?

Table 6: An instance from TFRKN

$[q_1, q_2, \cdots, q_n]$ and $[gt_1, gt_2, \cdots, gt_m]$.

$$Attr(w_i^{(l)}) = \overline{w}_i^{(l)} \int_{\beta=0}^{1} \frac{\mathrm{d}P_{[t_1,\cdots,t_n],y}(\beta\overline{w}_i^{(l)})}{\mathrm{d}w_i^{(l)}}\mathrm{d}\beta \tag{10}$$

$$\widetilde{\mathrm{Attr}}(query, w_i^{(l)}) = \frac{1}{m}\sum_{k=1}^{m}\overline{w}_{i,k}^{(l)}\int_{\beta=0}^{1}\frac{\mathrm{d}P_{[q_1,\cdots,q_n,\cdots,a_{k-1}],gt_k}(\beta\overline{w}_{i,k}^{(l)})}{\mathrm{d}w_{i,k}^{(l)}}\mathrm{d}\beta \tag{11}$$

where a_i represents the generated token with the highest predicted probability at i-th time. Due to the intractability of the continuous integration in Equation 10, an approximation is made using Riemann integration (equation 12). Substituting Equation 12 into Equation 11 yields Equation 13.

$$Attr(w_i^{(l)}) = \frac{\overline{w}_i^{(l)}}{N}\sum_{j=1}^{N}\frac{\partial P_{[t_1,\cdots,t_n],y}(\frac{j}{N}\overline{w}_i^{(l)})}{\partial w_i^{(l)}} \tag{12}$$

$$\widetilde{Attr}(query, w_i^{(l)}) = \frac{1}{m}\sum_{k=1}^{m}\frac{\overline{w}_{i,k}^{(l)}}{N}\sum_{j=1}^{N}\frac{\partial P_{[q_1,\cdots,q_n,a_1,\cdots,a_{k-1}],gt_k}(\frac{\overline{w}_{i,k}^{(l)}}{N})}{\partial w_{i,k}^{(l)}} \tag{13}$$

Given that knowledge neurons surpass linguistic expressions and govern the expression of authentic knowledge, we retain knowledge neurons shared by more than $p\%$ queries as Equation 14.

$$\begin{aligned} KN &= \bigcap_{k=1}^{L} KN_{query_k} \\ KN_{query_k} &= \{w_i^{(l)} \mid \widetilde{Attr}(query_k, w_i^{(l)}) > \tau, \forall i, l\} \end{aligned} \tag{14}$$

C Experimental Details

We present a comprehensive overview of our experimental setup.

Intersection of LLMs Experiments are conducted using a refined subset of TFRKN dataset. To ensure that LLMs know each factual element required by the factual reasoning questions, we meticulously filtered out unqualified data points for each model. By taking the intersection of these filtered datasets, we culled a dataset comprising 1072 qualified data points.

Indentification of KNs The process of identifying KNs for each fact triplet proves to be the most computationally intensive, with each model taking 96 GPU hours to find all KNs. In the context of the location experiment, we configured the integrated gradient steps to 20 and set the parameter of the shared percentage of coarse neurons to 0.2. The experiments were executed on a system equipped with NVIDIA A100 80GB GPUs, and further details of the software environment are available in our code repository. All experimental results are the mean values of three repetitive experiments.

D Construction of Contextual Conflict

Knowledge Distraction We manually constructed a set of irrelevant fact statements S. S does not involve any entities or relations in TFRKN to ensure "unrelated" property. Each two-hop question randomly selects a knowledge distraction from this set.

Knowledge Conflict We constructed contexts that conflict with the first-hop fact and that conflict with the second-hop fact for each two-hop question respectively. The method is as follows: we manually designed templates T for all relations involved in the TFRKN dataset. Assuming there is a fact (s, r, o), we collect the set of candidate objects related to r in the dataset, select an o^* that is not equal to o as the new fabricated fact (s, r, o^*), and apply the template of the relation in T to obtain the knowledge conflict context corresponding to (s, r, o).

E Additional Experimental Results

Non-overlap of Knowledge Neurons Based on the KNs identified in our experiments, we conducted a verification of non-overlap. To achieve this, we randomly sampled 6,000 pairs of distinct

	LLaMA2			**LLaMA3**			**Mistral**		
	Avg.	Med.	Max.	Avg.	Med.	Max.	Avg.	Med.	Max.
Number of KNs per fact	26.4	22.0	53.0	28.3	25.0	52.0	22.3	26.0	59.0
Number of pairwise intersections	2.7	2.0	6.0	3.7	3.0	4.0	1.0	0.0	4.0

Table 7: The KNs for different facts may vary significantly (Avg.: average, Med.: median, Max.: maximum).

Sentences ending with Paris	**LLaMA2-7B**	**LLaMA3-8B**	**Mistral-7B**
The capital of France is	**1.98**	**2.25**	**2.04**
The capital city of France is	**2.01**	**2.27**	**2.06**
The Louvre Museum is situated in the city of	1.10	1.46	1.55
The Seine River flows gracefully through the heart of	1.27	1.47	1.70
The City of Love refers to the city of	1.24	1.70	1.41
The City of Light refers to the city of	1.25	1.65	1.53
The capital The football club Paris Saint-Germain is based in the city of France is	1.24	1.40	1.66
The Eiffel Tower is one of the most iconic landmarks in the city of	1.37	1.80	1.64

Table 8: KN Scores corresponding to $(France, capital, Paris)$ for different sentences which end with Paris.

relational facts and calculated the number of intersecting KNs between each pair. The statistics of overlapping KNs are shown in Table 7.

Verification of Basic Assumptions We present compelling results from small-scale case studies, which prove that when LLMs predict the same word, the metric of KN Scores would be high only when the process involves fact retrieval. For illustration, we use $(France, capital, Paris)$ as an example, whose KNs cover 26 neurons. KN Scores are computed across these 26 neurons as the metric. Then we construct sentences that end with "Paris" and then replace "Paris" with a blank, prompting LLMs to predict the missing word. To ensure that the LLMs predict "Paris" as the final token, we design straightforward and commonsense sentences and verify that the LLMs would indeed predict "Paris" and then assess the knowledge-expressing prompts and compare them with non-knowledge-expressing prompts by analyzing their KN Scores.

"The capital of France is" and "The capital city of France is" are knowledge-expressing prompts, which consistently exhibit higher KN Scores compared to other examples, even though LLMs predict "Paris" for all these sentences. This experiment illustrates that KNs for $(France, capital, Paris)$ are activated mostly when LLMs recall $(France, capital, Paris)$, not when they make a specific predictive word "Paris".