Investigating Multi-Hop Factual Shortcuts in Knowledge Editing of Large Language Models

Anonymous ACL submission

Abstract

Recent work has showcased the powerful capability of large language models (LLMs) in recalling knowledge and reasoning. However, the reliability of LLMs in combining these two capabilities into reasoning through multi-hop facts has not been widely explored. This paper systematically investigates the possibilities for LLMs to utilize shortcuts based on direct connections between the initial and terminal entities of multi-hop knowledge. We first explore the existence of factual shortcuts through Knowledge Neurons, revealing that: (i) the strength of factual shortcuts is highly correlated with the frequency of co-occurrence of initial and terminal entities in the pre-training corpora; (ii) few-shot prompting leverage more shortcuts in answering multi-hop questions compared to chain-of-thought prompting. Then, we analyze the risks posed by factual shortcuts from the perspective of multi-hop knowledge editing. Analysis shows that approximately 20% of the failures are attributed to shortcuts, and the initial and terminal entities in these failure instances usually have higher co-occurrences in the pre-training corpus. Finally, we propose erasing shortcut neurons to mitigate the associated risks and find that this approach significantly reduces failures in multiple-hop knowledge editing caused by shortcuts. Code is publicly available at *Anonymous*.

1 Introduction

Large Language Models (LLMs) such as ChatGPT (OpenAI, 2022) and LLaMA-2 (Touvron et al., 2023), have impressive world knowledge modeling and reasoning capabilities within their parameters (Zhao et al., 2023; Hao et al., 2023). When leveraging these two capabilities, it is intuitively anticipated that LLMs should be capable of reliably answering multi-hop knowledge questions without any difficulty (Press et al., 2023).

Nonetheless, the underlying reasoning processes of LLMs in responding to multi-hop knowledge questions have not received thorough investigation. Ideally, an LLM would systematically derive each single-hop answer and culminate in the correct result. However, in reality, LLMs may leverage factual shortcuts learned from pre-training corpora to directly obtain the final answer without performing intermediate reasoning.

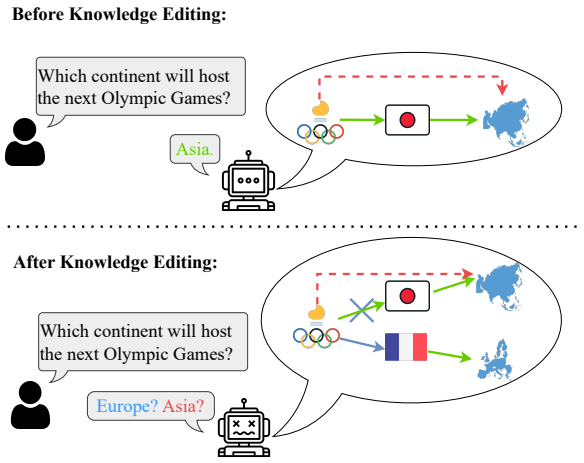


Figure 1: An illustrative example of a multi-hop factual shortcut in LLMs. The LLM may have directly encoded multi-hop knowledge (red) during the pre-training phase, which results in inconsistencies after a single-hop knowledge editing.

For conventional multi-hop question answering, the consistency of the final endpoints of shortcuts and multi-hop reasoning results may not cause risks and could even remain unnoticed. However, with the constant evolution of world knowledge, knowledge editing techniques are garnering increased attention (Wang et al., 2023b). After knowledge editing, factual shortcuts in multi-hop scenarios may cause significant inconsistency.

Figure 1 illustrates the potential pitfalls associated with factual shortcuts. During the pre-training phase, an LLM may have forged a direct association between the next Olympic Games and Asia. Consequently, when queried with the prompt:

"Which continent will host the next Olympic Games", the LLM might bypass the need for reasoning about the country and can directly furnish the correct answer. However, applying knowledge editing to the LLM, e.g., updating the host country of the Olympic Games to France, can expose a vulnerability. The persistence of the established shortcut may lead the LLM to consistently output "*Asia*" as the host continent even after the change, instead of the correct "*Europe*", thereby impeding the success of multi-hop knowledge editing.

In this paper, we systematically investigate the possibilities for LLMs to utilize factual shortcuts based on direct connections between the initial and terminal entities of multi-hop knowledge. Firstly, we rethink and formalize the process through which LLMs reason about multi-hop knowledge. **We introduce the hypothesis that LLMs may leverage factual shortcuts from pre-training corpora to facilitate cross-step reasoning**.

Then, we deeply explore the existence of factual shortcuts. We conduct a frequency analysis of co-occurrences between the initial subject and terminal object of multi-hop knowledge instances in pre-training corpora. Additionally, we employ Knowledge Neurons (Dai et al., 2022) to quantify the overlap between the activated neurons for multi-hop questions and all single-hop questions. A low degree of overlap suggests that the reasoning pattern of LLMs in response to multi-hop questions is inconsistent with that of single-hop questions, indicating the presence of shortcuts. Our experiments on multi-hop knowledge reveal that:

(i) Few-shot questions exhibit more shortcuts in comparison to chain-of-thought questions, suggesting that **LLMs often arrive at multi-hop knowledge answers using unexpected cross-step reasoning patterns**.

(ii) Knowledge instances with a higher co-occurrence frequency between initial subjects and terminal objects tend to have more shortcuts, indicating **a strong correlation between the existence of multi-hop factual shortcuts and the word frequencies learned by LLMs during pre-training phase**.

Additionally, to provide insights into the potential risks associated with multi-hop factual shortcuts, we conduct a detailed analysis of the reasons behind the failures in multi-hop knowledge editing. We find that **approximately 20% of the failure instances are attributed to multi-hop factual shortcuts**. Furthermore, **shortcut failure instances often exhibit higher co-occurrence frequencies of the initial and terminal entities**, providing compelling evidence that the presence of shortcuts may disrupt the multi-hop reasoning consistency of LLMs after knowledge editing.

Finally, we explore the feasibility of employing Knowledge Neurons to eliminate factual shortcuts. We erase crucial neurons associated with factual shortcuts that co-occurred more than 10 times in the pre-training corpus. Results show that **the failure rate of multiple-hop knowledge editing caused by shortcuts significantly decreased, leading to an overall improvement in the success rate after our erasing approach**. We hope this work can facilitate increased interest in exploring the multi-hop reasoning capabilities of LLMs and constrain reasoning shortcuts during the pre-training stage.

2 Rethinking the Multi-Hop Knowledge

A basic fact can be formulated as a single-hop knowledge tuple $t = (s, r, o)$ with a subject (s), a relation (r), and an object (o). For each query, we ask the LLM if the object is correct given the subject and the relation $\mathbb{1}\{f(T(s, r)) = o\}$, where f and T denote the outputs of the LLM and the prompt template for splicing s and o into a cloze-style form.

In this paper, we mainly focus on the multi-hop knowledge, which comprises a chain of single-hop knowledge:

$$\mathcal{T} = \langle (s_1, r_1, o_1), ..., (s_n, r_n, o_n) \rangle, \quad (1)$$

where $s_i = o_{i-1}$. For each query, we directly ask the LLM if the terminal object is correct given the initial subject and the chain relation $\mathbb{1}\{f(T(s_1, r_{mul})) = o_n\}$, where $r_{mul} = r_1 \rightarrow ... \rightarrow r_n$. This question can also be formulated as asking the LLM of the knowledge tuple $t_{mul} = (s_1, r_{mul}, o_n)$, which proves unproblematic in general multi-hop question-answering, as t_{mul} and $\mathcal{T}$ share the same endpoint o_n.

However, t_{mul} is in fact a shortcut, treating a chain of relations as a separate composite relation. If a knowledge-editing approach is employed to modify the intermediate entity o_i to o_i^*, the final answer of $\mathcal{T}$ will be altered. Since t_{mul} overlooks the intermediate entity, its answer remains unaffected by knowledge editing.

Taking the multi-hop question of "*Which continent will host the next Olympic Games*" as an illustrative example, if we edit the knowledge of the

"*country*" from Japan to France, according to the chain-relation reasoning, the "*continent*" hosting the Olympic Games should be converted to Europe. However, if a composite relation is employed, the "*continent*" would remain unchanged despite alterations in the "*country*".

A causal LLM probably encodes such composite knowledge during the pre-training phase. The initial subject s_1 and the terminal object o_n are likely to have direct associations in the corpus. Still taking the example above, an LLM may have learned the knowledge *(the next Olympic Games, continent of the country, Asia)* from the corpus directly, neglecting the causal relationship between the country and the continent to which it belongs. Therefore, for multi-hop knowledge, LLMs may potentially arrive at the correct answer through step-wise reasoning, but it is more likely that they memorize the outdated answer by leveraging the co-occurrence relationships in the pre-training corpus.

3 Exploring the Existence of Factual Shortcuts

In this section, we explore the extent of shortcuts in multi-hop question-answering. Concretely, we first validate the correlation between multi-hop shortcuts and the word frequency in the pre-training corpus. Then, we locate crucial neurons in single-hop, few-shot, and chain-of-thought question-answering tasks to further elucidate the degree of potential factual shortcuts.

3.1 Probing Shortcuts in Pre-training Corpus

Our analysis centers specifically on the MQUAKE-CF-3K dataset released by Zhong et al. (2023). It comprises 3,000 instances of multi-hop question-answering for knowledge editing extracted from Wikidata (Vrandecic and Krötzsch, 2014), which can be adopted in subsequent sections for further investigating potential risks introduced by these multi-hop factual shortcuts (details are shown in Appendix A).

Considering that the existence of factual shortcuts may drive from pre-training corpora, we first compute the frequency of co-occurrence of the initial subject s_1 and the terminal object o_n among these 3,000 items of knowledge on Wikipedia (20231101-en) corpus, which contains approximately 6.41M rows of texts. We choose this corpus due to its comprehensive coverage of global knowledge and its frequent utilization as a significant component in the pre-training corpora for most LLMs. If s_1 and o_n co-occur within the same paragraph, it is highly plausible that the LLM establishes a direct connection between them during the pre-training phase.

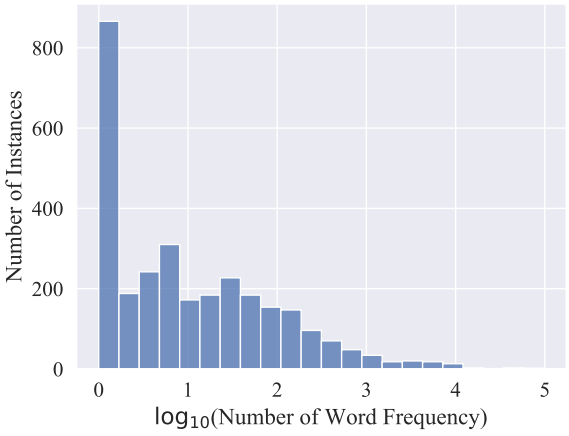


Figure 2: Frequency analysis of multi-hop knowledge shortcuts in Wikipedia.

We conduct a frequency analysis of the occurrences of these multi-hop knowledge shortcuts in the corpus (Figure 2). It can be observed that more than 2/3 of instances exhibited various degrees of shortcuts, with some even appearing over 10,000 times. This indicates that **certain pieces of knowledge exhibit significant multi-hop shortcuts**, which could potentially influence the reasoning processes of LLMs.

Moreover, we select several examples with high and low frequencies for illustration (Table 1). It can be observed that instances with high frequency exhibit a clear direct connection between s_1 and o_n. For instance, "*Twitter*" is inherently strongly associated with "*the United States*", obviating the need to think about the country of citizenship of "*Twitter's CEO*". In contrast, there is no apparent connection between "*Jerry Rivers*" and "*Donald Trump*", necessitating the prior derivation of the nationality of "*Jerry Rivers*" to arrive at the correct answer. Since "*Jerry Rivers*" and "*Donald Trump*" rarely co-occur in the pre-training corpus, LLMs may not contain factual shortcuts related to such multi-hop knowledge.

3.2 Quantifying Shortcuts Using Knowledge Neurons

Methods. The presence of multi-hop factual shortcuts may result in a divergence in the reasoning mechanisms employed by the LLM when

Subject (s_1)	Object (o_n)	Multi-Hop Question	Frequency
Rhode Island	English	Which languages are spoken, written, or signed in Rhode Island as the head of government there?	42754
Twitter	United States of America	What is the country of citizenship of Twitter's CEO?	35435
Fanta	Atlanta	What is the location of the headquarters of the manufacturer of Fanta?	25834
Jerry Rivers	Donald Trump	Who is the head of state of the country whose citizen is Jerry Rivers?	0
Alvar Aalto	Mikael Agricola	Who is the creator of the content in the language or languages spoken by Alvar Aalto?	0
Nick Rimando	London	What is the capital of the country where the sport of Nick Rimando's position is originated?	0

Table 1: Examples of multi-hop knowledge with high and low frequency of co-occurrence of s_1 and o_n.

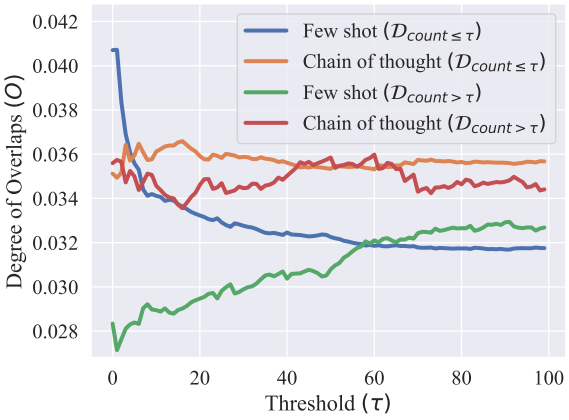


Figure 3: The degree of overlaps employed by GPT-J in handling multi-hop questions and all single-hop questions with varying word frequencies in pre-training corpora under few-shot prompts and chain-of-thought prompts. It is expected that the instances from $\mathcal{D}_{\text{count}>\tau}$ contain more potential factual shortcuts.

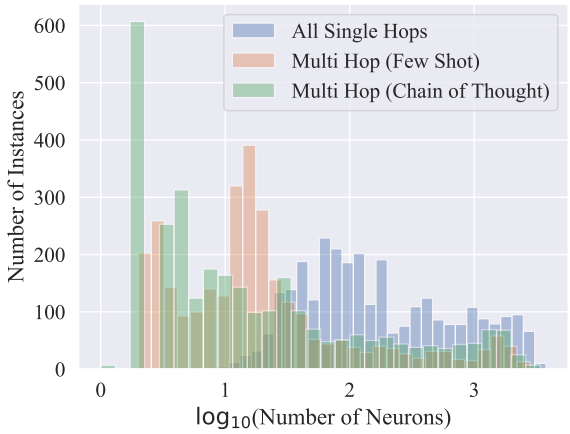


Figure 4: Distribution of the number of activated neurons in GPT-J across different questions.

responding to multi-hop questions as opposed to directly answering individual single-hop questions. To quantify the disparities, we employ Knowledge Neurons (KN) proposed by Dai et al. (2022) to locate crucial neurons activated by the LLMs when responding to various questions. Specifically, it gradually changes each neuron $w_i^{(l)}$ stored in FFN from 0 to its original value $\overline{w}_i^{(l)}$ and meanwhile integrates the gradients. We use the Riemann approximation as a substitution for continuous integrals:

$$\tilde{\text{Attr}}(w_i^{(l)}) = \frac{\overline{w}_i^{(l)}}{m} \sum_{k=1}^{m} \frac{\partial P(\frac{k}{m}\overline{w}_i^{(l)})}{\partial w_i^{(l)}}, \quad (2)$$

where $P(w_i^{(l)}) = p(y|x, w_i^{(l)} = \hat{w}_i^{(l)})$ is the probability of the correct answer predicted by the LLM when changing the value of neuron $w_i^{(l)}$ to $\hat{w}_i^{(l)}$, and m is the number of the approximation steps. We choose neurons with attribution values larger than v as crucial neurons reflecting LLM decision-making patterns:

$$\mathbf{N} = \left\{ w_i^{(l)} | \text{Attr}(w_i^{(l)}) > v \right\}. \quad (3)$$

In this paper, we set m to 20 and the attribution threshold v to 0.2. In the scenario of a multi-hop question devoid of any shortcuts, it should ideally encompass a broader array of crucial neurons inherent to single-hop questions, except those specifically dedicated to lower-level components such as lexical and syntactic neurons. Hence, we define O as the degree of overlap between the reasoning patterns of multi-hop knowledge answers and all single-hop knowledge answers:

$$O = \frac{|\mathbf{N}_{\mathcal{T}} \cap \mathbf{N}_{t_{mul}}|}{|\mathbf{N}_{\mathcal{T}}|}, \quad (4)$$

where $\mathbf{N}_{\mathcal{T}}$ denotes the intersection of crucial neurons for all single-hop questions, $N_{t_{mul}}$ denotes the set of crucial neurons for multi-hop questions. A higher degree of overlap indicates that LLM's reasoning patterns for answering multi-hop questions are more closely aligned with those for answering single-hop questions.

It is noteworthy to emphasize that our objective does not entail the precise localization of neurons storing knowledge; rather, we aim to discern the decision-making processes of the LLMs across various questions. Despite Anonymous (2024)'s skepticism regarding whether neurons uncovered by KN in the FFN truly constitute "knowledge",

these neurons may store intricate "token expression patterns" that can still elucidate the LLM's decision-making processes.

We separately evaluate the degree of shortcuts in few-shot and chain-of-thought multi-hop questions. All single-hop questions and few-shot multi-hop questions utilize the same demonstrations, while chain-of-thought multi-hop questions employ prompts with similar semantics (details provided in Appendix B). Besides, we partition the original dataset $\mathcal{D}_o$ into two subsets $\mathcal{D}_{\text{count}\leq\tau}$ and $\mathcal{D}_{\text{count}>\tau}$ based on word frequencies, where τ represents the threshold for word frequencies. We compute the degree of shortcuts for GPT-J (Figure 3).

Main Results. It can be observed that **the LLM adheres to a greater extent to reasoning patterns overlapping with those for single-hop questions under the chain-of-thought prompt**. This observation suggests that the chain-of-thought prompt indeed serves to induce the LLM to engage in step-wise reasoning. It also aligns with our hypothesis that LLMs tend to prioritize the utilization of latent multi-hop factual shortcuts, relinquishing them only when explicitly prompted to engage in step-wise reasoning. Furthermore, the instances within $\mathcal{D}_{\text{count}>\tau}$ exhibit lower degrees of reasoning overlap, suggesting that **LLMs indeed learn the shortcut associations between s_1 and o_n, with word frequencies significantly influencing the strength of these shortcuts**.

Interestingly, although the overlap rates vary across different scenarios, their values remain low. We analyze the distribution of the number of activated knowledge neurons for different instances (Figure 4). Since single-hop knowledge typically involves 2-4 questions, the number of activated neurons is an order of magnitude higher than that for multi-hop questions. Moreover, activated neurons, in addition to reflecting the inherent knowledge, may also be influenced by factors such as the lexical and syntactic aspects of sentences. Hence, the reasoning overlap rates tend to be maintained at a low value.

4 Exploring the Potential Risks of Factual Shortcuts

While these shortcuts may not have a significant impact on the results in general multi-hop question answering, their potential risks can be magnified in the context of knowledge editing. Zhong et al. (2023) have observed poor performance of LLMs in multi-hop knowledge editing. In this section, we will specifically analyze the reasons for the failure of multi-hop knowledge editing, particularly under the influence of shortcuts.

Concretely, we employ various knowledge editing methods to modify single-hop knowledge instances in MQUAKE-CF-3K and pose three different multi-hop questions about the edited knowledge. Subsequently, we quantify the effects of various knowledge editing methods and categorize error instances into three categories.

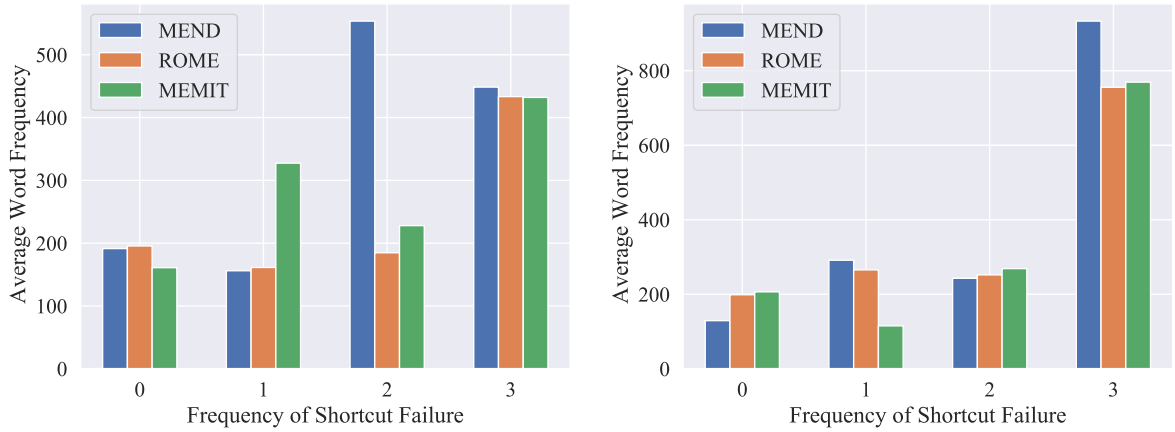
Failure Categories. We consider three key categories of failures. The first category of failure stems from the unsuccessful editing of single-hop knowledge. We designate the set of failures in this category as $\mathcal{F}_{\text{single}}$. The second and third categories are built upon the assumption of successfully editing all single-hop knowledge instances, yet the LLM still fails to answer multi-hop questions correctly. The second category signifies cases where the answer to multi-hop knowledge questions remains the same as the original unedited answer. We denote this set as $\mathcal{F}_{\text{shortcut}}$. Given that all single-hop questions can be answered correctly, the persistence of the original result in multi-hop questions indicates the existence of shortcuts. The third category involves the LLM providing alternative incorrect answers, potentially arising from hallucinations or other reasons. We denote this set as $\mathcal{F}_{\text{other}}$.

For each multi-hop edited knowledge, we interrogate the LLM with three distinct multi-hop questions. All multi-hop questions are prefixed with the same few-shot template comprising 16 demonstrations, which is consistent with the setup of Zhong et al. (2023). We calculate the percentage of editing successes ($\mathcal{S}$) and failures ($\mathcal{F}$) within three questions. Detailed experimental settings can be seen in the Appendix C.

Main Results. Table 2 presents the analysis results. Consistent with the findings of Zhong et al. (2023), knowledge editing algorithms exhibit catastrophic failures when addressing multi-hop factual questions, with only approximately 10%-20% of instances avoiding complete errors across three queries. $\mathcal{F}_{\text{single}}$ stems from the editing failure of LLMs in addressing single-hop questions. Since multi-hop questions may necessitate more than one edit, it may be slightly higher than the edit-wise failure rate. $\mathcal{F}_{\text{other}}$ may originate from the insufficient reasoning capabilities of LLMs or the

		$\mathcal{S}$				$\mathcal{F}_{\text{single}}$	$\mathcal{F}_{\textbf{shortcut}}$				$\mathcal{F}_{\text{other}}$			
		$i=1$	$i=2$	$i=3$	Sum	Sum	**i = 1**	**i = 2**	**i = 3**	**Sum**	$i=1$	$i=2$	$i=3$	Sum
GPT-J (6B)	MEND	4.27	4.53	14.17	22.97	33.03	**3.93**	**3.17**	**11.87**	**18.97**	5.40	4.97	33.47	43.84
	ROME	2.07	2.30	4.57	8.94	39.87	**3.13**	**2.27**	**9.17**	**14.57**	3.17	3.63	41.13	47.93
	MEMIT	2.17	1.97	4.87	9.01	33.37	**4.10**	**3.63**	**11.47**	**19.20**	4.20	4.00	43.27	51.47
LLaMA-2 (7B)	MEND	7.40	4.80	7.77	19.97	43.57	**5.63**	**5.70**	**9.63**	**20.96**	5.90	6.20	26.90	39.00
	ROME	5.33	3.00	3.83	12.16	25.37	**6.30**	**6.17**	**11.67**	**24.14**	6.80	7.00	44.67	58.47
	MEMIT	5.13	3.60	3.83	12.56	32.00	**6.00**	**5.47**	**10.17**	**21.64**	6.20	7.13	40.33	53.66

Table 2: The percentage of successful ($\mathcal{S}$) and failed ($\mathcal{F}$) multi-hop knowledge edits, where i denotes the frequency of success or failure within the three queries, "Sum" denotes the cases with at least one success or failure. We mainly focus on failures caused by factual shortcuts ($\mathcal{F}_{\text{shortcut}}$).



(a) GPT-J (6B) (b) LLaMA-2 (7B)

Figure 5: The average co-occurrence frequency of s_1 and o_n in the pre-training corpus. The horizontal axis represents the number of occurrences of shortcut failures across three queries.

hallucinations generated during editing. While we utilize few-shot prompts instead of chain-of-thought prompts to expose factual shortcuts, it may also increase $\mathcal{F}_{\text{other}}$.

It is noteworthy that $\mathcal{F}_{\text{shortcut}}$ also constitutes a significant proportion. This type of failure implies that LLMs respond with old ground truth for multi-hop questions while capable of correctly answering single-hop questions after all edits. In other words, shortcuts enable LLMs to conveniently utilize the r_{mul} hard-coded during the pre-training phase to directly obtain results, without genuinely engaging in multi-hop knowledge reasoning. **Experiments indicate that these factual shortcuts are prevalent across various knowledge types in LLMs**.

To further investigate the connection between shortcut failures and falsely learned relations in the pre-training corpus, we analyze the relationship between the average co-occurrence frequency of entities and the occurrence frequency of shortcut failures (Figure 5). We observe that **instances with higher occurrences of shortcut failures, particularly those with three failures, exhibit higher word co-occurrence frequencies between s_1 and o_n.** This suggests that LLMs are highly likely to leverage the multi-hop knowledge hardcoded during the pre-training phase as reasoning shortcuts. The presence of these factual shortcuts significantly diminishes the reliability and plausibility of LLMs' reasoning. In the context of multi-hop knowledge editing, the LLMs are easily entangled in the confusion between old shortcut knowledge and new multi-hop knowledge.

5 Reducing Multi-Hop Factual Shortcuts

The existence of multi-hop factual shortcuts reveals the unreliability of current LLMs' reasoning and increases the risk of failures in multi-hop knowledge editing. Since these shortcuts represent knowledge hardcoded into LLMs during the pre-training phase, it is challenging to eliminate these factual shortcuts fundamentally unless there are substantial changes in the pre-training phase.

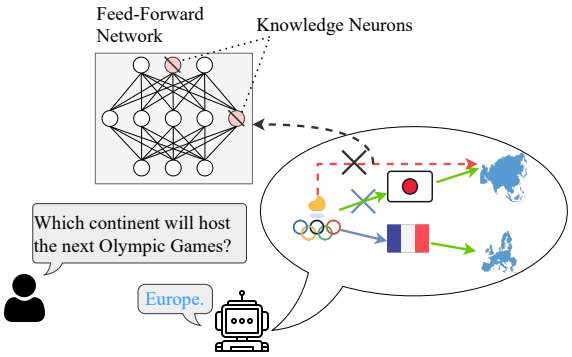


Figure 6: An illustrative example for reducing multi-hop factual shortcuts.

Methods. To reduce the risks of multi-hop factual shortcuts and further validate the hypotheses presented in this paper, we adopt a simple yet effective method inspired by Dai et al. (2022) to erase these shortcuts (Figure 6). Compared to Figure 1, we erase crucial neurons related to the red factual shortcuts, compelling the LLM to answer the continent that will host the next Olympic Games using the correct path of reasoning after knowledge editing.

Specifically, we use the integral gradient algorithm to locate the crucial neurons associated with multi-hop knowledge questions and set them to zero. For each piece of multi-hop knowledge, we query with three questions to obtain the intersection of crucial neurons. Based on the previous experiments (Figure 3), we posit that multi-hop knowledge with a co-occurrence frequency exceeding 10 exhibits evident shortcuts. Consequently, we proceeded to eliminate these multi-hop factual shortcuts from the dataset $\mathcal{D}_{\text{count}>10}$. We compute the percentage of editing success ($\mathcal{S}$) and shortcut failure rate ($\mathcal{F}_{\text{shortcut}}$) for multi-hop knowledge editing before and after the erase of factual shortcuts, respectively.

Main Results. Table 3 presents the success rate and shortcut failure rate of multi-hop knowledge editing before and after the erase of factual shortcuts on $\mathcal{D}_{\text{count}>10}$. Compared to Tabel 2, both the success rate and shortcut failure rate of multiple-hop knowledge editing have increased on $\mathcal{D}_{\text{count}>10}$. The result implies that instances with factual shortcuts are inherently more amenable to editing, yet the presence of factual shortcuts also entails a higher level of risk for these instances. Thus, these latent factual shortcuts are far more harmful than we realize.

Furthermore, **the erasing of shortcuts can significantly reduce the risks associated with shortcut failures, leading to an appreciable improvement in the success rate of multi-hop knowledge editing**. Due to the incapacity of knowledge editing methods to address shortcut knowledge t_{mul}, inconsistencies arise in LLMs' reasoning results. By erasing neurons corresponding to t_{mul}, we ensure that LLMs reason along the correct path, thereby enhancing the success rate.

However, despite the efficacy of this approach in mitigating the risk posed by factual shortcuts to specific knowledge, it cannot serve as a comprehensive solution to the problem. Due to the ubiquitous nature of such shortcuts, it is impractical to review and erase crucial neurons for every multi-hop knowledge. Fundamentally, to attain a trustworthy LLM with genuine multi-hop reasoning capabilities, it is imperative to address the issue at the pre-training stage to explore improved pre-training methodologies.

6 Related Work

In this section, we discuss two lines of research that are key to our work: knowledge editing and multi-hop reasoning.

6.1 Knowledge Editing

Numerous studies have explored efficient knowledge editing methods for LLMs, seeking resolutions to challenges arising from outdated knowledge. One prevalent and intuitive approach involves employing external memorization, wherein new knowledge is incorporated through external context or parameters, without necessitating modifications to the LLM weights (Mitchell et al., 2022b; Dong et al., 2022; Huang et al., 2023; Zheng et al., 2023; Zhong et al., 2023). While these approaches are simple and effective in ensuring consistency, the substantial influx of supplementary knowledge may result in redundancy and low timeliness at a later stage (Wang et al., 2023b).

Another line of work focuses on directly updating the LLM parameters. Some investigations are dedicated to constrained fine-tuning (Chen et al., 2020; Lee et al., 2022) or meta-learning (Lee et al., 2022; Mitchell et al., 2022a), which update the full parameters of LLMs. The other investigations involve a preliminary stage of knowledge localization before editing, premised on the assumption that knowledge is stored in

			$\mathcal{S}\uparrow$				$\mathcal{F}_{\text{shortcut}}\downarrow$			
			$i=1$	$i=2$	$i=3$	Sum	$i=1$	$i=2$	$i=3$	Sum
GPT-J (6B)	MEND	Before Erasing	4.46	5.13	**19.56**	29.15	**4.08**	**4.18**	17.57	25.83
		After Erasing	**5.79**	**5.41**	18.42	**29.62**	4.75	4.84	**15.76**	**25.35**
	ROME	Before Erasing	2.09	**2.94**	**8.64**	13.67	2.75	**3.23**	12.82	18.80
		After Erasing	**4.47**	**2.94**	8.36	**15.77**	**2.57**	3.33	**11.62**	**17.52**
	MEMIT	Before Erasing	1.61	**2.94**	7.98	12.53	4.27	5.60	16.05	25.92
		After Erasing	**3.32**	2.66	**8.07**	**14.05**	**3.23**	**4.65**	**14.25**	**22.13**
LLaMA-2 (7B)	MEND	Before Erasing	**9.21**	5.79	**9.97**	**24.97**	**6.27**	**7.03**	17.76	31.06
		After Erasing	8.36	**6.08**	9.31	23.75	7.79	8.17	**5.51**	**21.47**
	ROME	Before Erasing	5.98	4.65	**7.03**	17.66	**6.84**	6.93	18.33	32.10
		After Erasing	**7.50**	**4.75**	6.93	**19.18**	7.03	**6.36**	**11.97**	**25.36**
	MEMIT	Before Erasing	5.60	4.84	**7.12**	17.46	**6.08**	6.17	17.09	29.34
		After Erasing	**8.36**	**5.03**	6.74	**20.13**	6.36	**5.88**	**9.88**	**22.12**

Table 3: Success rate and shortcut failure rate of multi-hop knowledge editing before and after the erase of factual shortcuts on $\mathcal{D}_{\text{count}>10}$.

the form of key-value memories within the two-layer Feedforward Neural Network (FFN) (Geva et al., 2021). Dai et al. (2022) located and refined knowledge neurons (KN) through integral gradients (Sundararajan et al., 2017). Meng et al. (2022) et al. proposed the Rank-One Model method (ROME) to insert new knowledge in a specific FFN layer, while MEMIT (Meng et al., 2023) further extended address scenarios of mass editing.

While the effectiveness of single-hop knowledge editing has been thoroughly investigated, there is a notable dearth of attention given to multi-hop knowledge editing. Zhong et al. (2023) systematically focused on this issue by introducing the multi-hop knowledge editing evaluation benchmarks MQUAKE-CF and MQUAKE-T. Their findings revealed catastrophic performance degradation of existing knowledge editing methods. In this paper, we further investigate and elucidate the repercussions stemming from the presence of reasoning shortcuts in multi-hop knowledge editing.

6.2 Multi-Hop Reasoning

Multi-hop reasoning is often seen as a weakness for LLMs (Huang and Chang, 2023). Early efforts commonly employed in-context prompting, which involves the provision of few input-output demonstrations to LLMs (Brown et al., 2020; Zhao et al., 2021; Chen et al., 2022). This approach enables LLMs to solve problems through reasoning implicitly. However, its effectiveness diminishes significantly when confronted with multi-hop questions (Valmeekam et al., 2022). To incentivize LLMs to engage in explicit multi-hop reasoning, the concept of *chain-of-thought* was introduced by Wei et al. (2022). It encourages the LLM to think step by step and output intermediate deductive steps (Chu et al., 2023). In this paper, we elucidate the process by which LLMs handle multi-hop question-answering from the perspective of factual shortcuts. We provide evidence that the chain-of-thought prompting compels LLMs to attend to the single-hop knowledge more faithfully.

7 Conclusion

In this paper, we systematically explore the latent factual shortcuts that LLMs may employ when answering multi-hop knowledge questions. We first demonstrate the strong correlation between the strength of factual shortcuts and the co-occurrence of the initial subject and the terminal object in pre-training corpora. Then, we delve into the potential risks introduced by these shortcuts in the context of multi-hop knowledge editing. Our exploration reveals that approximately 20% of failures can be attributed to factual shortcuts, particularly in instances characterized by high co-occurrences within pre-training corpora. Finally, we propose a straightforward yet efficient approach to mitigate shortcut failures in multi-hop knowledge editing by selectively erasing shortcut neurons. We advocate for increased research efforts directed towards exploring the true boundaries of LLMs in the realm of multi-hop reasoning, emphasizing the need to better constrain shortcut generation during the pre-training phase.

8 Limitations

We posit that Wikipedia serves as a comprehensive repository of global knowledge, thus making it a suitable substitute for the entirety of the pre-training corpora. However, despite our exhaustive traversal of the Wikipedia dataset to calculate the co-occurrence frequencies of initial and terminal entities, it is noteworthy that the pre-training corpora for LLMs often extend beyond the confines of this dataset. This potential discrepancy may introduce inaccuracies in statistical outcomes. We advocate for future investigations to extend statistical analyses to more expansive corpora.

For the erasing of factual shortcuts, our primary objective is to further substantiate the potential risks associated with these shortcuts, and the observed improvement in editing success rates after erasing serves to support this assertion. However, it is imperative to recognize that this approach functions as a mitigative measure, as the complete eradication of factual shortcuts through post-hoc removal is unattainable. A genuine and thorough elimination of factual shortcuts must be initiated during the pre-training phase, involving the alignment of LLMs' multi-hop reasoning capabilities with human-level proficiency.

Finally, due to space and resource constraints, we only conduct detailed experiments on GPT-J (6B) and LLaMA-2 (7B) and do not encompass all publicly accessible LLMs, such as PaLM (Chowdhery et al., 2022), OPT (Zhang et al., 2022), and Pythia (Biderman et al., 2023). We encourage future research to undertake comprehensive experiments on a broader spectrum of LLMs.

9 Ethical Statement

We conduct a reassessment of the multi-hop reasoning capabilities of LLMs and demonstrate that the presence of factual shortcuts may compromise the consistency of results in multi-hop knowledge editing. Since the approach itself is unbiased and all experiments are conducted on publicly available datasets, we believe that our work creates no potential ethical risk. Additionally, all use of existing artifacts is consistent with their intended use in this paper.

However, we have exposed the indiscriminate use of shortcuts by LLMs during multi-hop reasoning, raising concerns regarding their genuine reasoning capabilities. LLMs struggle to engage in step-wise reasoning akin to human cognitive processes, and the potential for parameter confusion may arise following the assimilation of new knowledge. These factors contribute to our perplexity concerning the black-box nature of LLMs and apprehensions regarding their application in security-sensitive domains. We advocate for more rigorous ethical scrutiny and improvements in LLMs to ensure alignment with the human reasoning process.

References

Anonymous. 2024. What does the knowledge neuron thesis have to do with knowledge? In *The Twelfth International Conference on Learning Representations*.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 2397–2430. PMLR.

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

Mingda Chen, Jingfei Du, Ramakanth Pasunuru, Todor Mihaylov, Srini Iyer, Veselin Stoyanov, and Zornitsa Kozareva. 2022. Improving in-context few-shot learning via self-supervised training. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 3558–3573. Association for Computational Linguistics.

Sanyuan Chen, Yutai Hou, Yiming Cui, Wanxiang Che, Ting Liu, and Xiangzhan Yu. 2020. Recall and learn: Fine-tuning deep pretrained language models with less forgetting. In *Proceedings of the 2020 Conference on Empirical Methods in Natural*

Language Processing, EMNLP 2020, Online, November 16-20, 2020, pages 7870–7881. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2023. A survey of chain of thought reasoning: Advances, frontiers and future. *CoRR*, abs/2309.15402.

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), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8493–8502. Association for Computational Linguistics.

Qingxiu Dong, Damai Dai, Yifan Song, Jingjing Xu, Zhifang Sui, and Lei Li. 2022. Calibrating factual knowledge in pretrained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5937–5947. 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, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 5484–5495. Association for Computational Linguistics.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8154–8173, Singapore. Association for Computational Linguistics.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1049–1065. Association for Computational Linguistics.

Zeyu Huang, Yikang Shen, Xiaofeng Zhang, Jie Zhou, Wenge Rong, and Zhang Xiong. 2023. Transformer-patcher: One mistake worth one neuron. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Kyungjae Lee, Wookje Han, Seung-won Hwang, Hwaran Lee, Joonsuk Park, and Sang-Woo Lee. 2022. Plug-and-play adaptation for continuously-updated QA. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 438–447. Association for Computational Linguistics.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), Vancouver, Canada, August 3-4, 2017*, pages 333–342. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Kevin Meng, Arnab Sen Sharma, Alex J. Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. 2022a. Fast model editing at scale. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. 2022b. Memory-based model editing at scale. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 15817–15831. PMLR.

OpenAI. 2022. Introducing chatgpt. `https://openai.com/blog/chatgpt`.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language

models. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 5687–5711. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Karthik Valmeekam, Alberto Olmo Hernandez, Sarath Sreedharan, and Subbarao Kambhampati. 2022. Large language models still can't plan (A benchmark for llms on planning and reasoning about change). *CoRR*, abs/2206.10498.

Denny Vrandecic and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

Peng Wang, Ningyu Zhang, Xin Xie, Yunzhi Yao, Bozhong Tian, Mengru Wang, Zekun Xi, Siyuan Cheng, Kangwei Liu, Guozhou Zheng, et al. 2023a. Easyedit: An easy-to-use knowledge editing framework for large language models. *arXiv preprint arXiv:2308.07269*.

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. 2023b. Knowledge editing for large language models: A survey. *CoRR*, abs/2310.16218.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *CoRR*, abs/2303.18223.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. 2023. Can we edit factual knowledge by in-context learning? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 4862–4876. Association for Computational Linguistics.

Zexuan Zhong, Zhengxuan Wu, Christopher D. Manning, Christopher Potts, and Danqi Chen. 2023. Mquake: Assessing knowledge editing in language models via multi-hop questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 15686–15702. Association for Computational Linguistics.

A Dataset

We select the MQUAKE-CF-3K dataset as the primary focus for exploration in this paper. It comprises 3,000 multi-hop English knowledge questions extracted from Wikipedia along with a corresponding knowledge editing task. We present essential information for one sample from the dataset (Table 4). For Section 3, we compute the crucial neurons of the first question within the 'questions' key, alongside the entirety of questions within the 'single_hops' key. For Section 4, we adopt knowledge editing methods of all knowledge encapsulated within the 'requested_rewrite' key. Furthermore, we augment the original dataset by introducing a new key, labeled as 'shortcut_frequency', which denotes the frequency of co-occurrence in the pre-training corpus between the initial subject and the terminal object for each instance. We commit to open-sourcing the dataset for subsequent research use.

B Prompts for Knowledge Neurons

We employ prompt templates similar to that utilized by Zhong et al. (2023) for finding crucial neurons. Given the substantial computational overhead associated with Knowledge Neurons, we adopt a 2-shot prompt, which is already sufficient for the LLM to comprehend the task and furnish accurate responses.

For all single-hop questions, we adopt the few-shot prompt shown in Table 5. Subsequently, we locate crucial neurons based on the probability of correct answer output by the LLM following the "A:" prefix.

For multi-hop questions, we adopt both the few-shot and chain-of-thought prompts. The few-shot prompt mirrors that of single-hop questions, while the chain-of-thought prompt is constructed with semantically approximate expressions. We require the LLM to articulate its reasoning process upon receiving the question. Then we locate crucial neurons based on the probability of correct answer output by the LLM following the "Answer:" prefix (see Table 6).

C Experimental Details

C.1 Language Models

Our experiments are conducted on GPT-J (6B) (Wang and Komatsuzaki, 2021) and LLaMA-2 (7B) (Touvron et al., 2023). The selection of GPT-J is motivated by the alignment with the pre-existing work on knowledge editing (Meng et al., 2022, 2023; Zhong et al., 2023), while opting for LLaMA-2 is motivated by its status as a recent, prominent open-source LLM representative, providing a robust reflection of the current capabilities of LLMs. We use the huggingface package (Wolf et al., 2020) for the specific implementation.

C.2 Knowledge Editing

We use the cloze-style statement templates for knowledge editing, which is consistent with the previous studies. We employ the EasyEdit package (Wang et al., 2023a) for the specific implementation. All licenses of these packages allow us for normal research use. The detailed specifics of the three knowledge editing methods that are employed in our training are as follows.

MEND. MEND (Mitchell et al., 2022a) trains a lightweight model editor network to produce edits to the LLM's weight when provided with the standard fine-tuning gradient. We train our editor network on the ZsRE dataset (Levy et al., 2017) with a maximum number of training steps of 100,000. We set the learning rate scale to be 1.0 during inference. All experiments edit the MLP weights in the last 3 Transformer blocks.

ROME. ROME (Meng et al., 2022) stands out as a popular method for knowledge localization and editing. It introduces a based on corruption and restoration to identify relevant layers storing knowledge. Subsequently, it inserts new knowledge by key selection and value optimization in the corresponding feed-forward network (FFN) layer. We perform the intervention at layer 5 for GPT-J (6B) and 6 for LLaMA-2 (7B). We compute the second-order momentum statistics using 100,000 examples of Wikitext in `fp32`. For the remaining hyperparameters, we adopt the default values specified in Meng et al. (2022).

MEMIT. MEMIT (Meng et al., 2023) is a subsequent work to ROME, designed to handle extensive knowledge edits. In this paper, we perform the intervention at layer {3, 4, 5, 6} for GPT-J (6B) and {4, 5, 6, 7} for LLaMA-2 (7B). We also compute the covariance statistics using 100,000 examples of Wikitext in `fp32`. For the remaining hyperparameters, we adopt the default values specified in Meng et al. (2023).

```
case_id: 16
requested_rewrite: [
    {
        prompt: {} is a citizen of
        target_new: Latvia,
        target_true: United States of America,
        subject: Jack Dorsey,
        question: What is the country of citizenship of Jack Dorsey?
    }
]
questions: [
    What is the country of citizenship of Twitter's CEO?
    From which country does Twitter's CEO hold citizenship?
    Which country's citizenship is held by the CEO of Twitter?
]
answer: United States of America
answer_alias: ...
new_answer: Latvia
new_answer_alias: ...
shortcut_frequency: 35435
single_hops: [
    {
        question: Who is the chief executive officer of Twitter?
        cloze: The chief executive officer of Twitter is
        answer: Jack Dorsey
        answer_alias: ...
    }
    {
        question: What is the country of citizenship of Jack Dorsey?
        cloze: Jack Dorsey is a citizen of
        answer: United States of America
        answer_alias: ...
    }
]
new_single_hops: [
    {
        question: Who is the chief executive officer of Twitter?
        cloze: The chief executive officer of Twitter is
        answer: Jack Dorsey
        answer_alias: ...
    }
    {
        question: What is the country of citizenship of Jack Dorsey?
        cloze: Jack Dorsey is a citizen of
        answer: Latvia
        answer_alias: ...
    }
]
```

Table 4: Critical information for a sample in the multi-hop knowledge editing dataset MQUAKE-CF-3K. We have added the 'shortcut_frequency' key to the original dataset to store the frequency of shortcuts appearing in Wikipedia.

Q: Who is the spouse of the US president? A: Jill Biden Q: In which country is the company that created Nissan 200SX located? A: Japan Q: [Input Question] A: [Output Answer]

Table 5: The few-shot prompt for Knowledge Neurons.

Question: Who is the spouse of the US president? Thoughts: The US president is Joe Biden. The spouse of Joe Biden is Jill Biden. Answer: Jill Biden. Question: In which country is the company that created Nissan 200SX located? Thoughts: Nissan 200SX was created by Nissan. Nissan is located in the country of Japan. Answer: Japan. Question: [Input Question] Thoughts: [Output Thoughts] Answer: [Output Answer]

Table 6: The chain-of-thought prompt for Knowledge Neurons.

C.3 Computational Budget

For all the experiments mentioned in this paper, we use one Nvidia A100-SXM4 GPU with 80GB memory. We spend about 100, 200, and 250 GPU hours exploring the existence of factual shortcuts (Section 3), exploring the potential risks of factual shortcuts (Section 4) and reducing multi-hop factual shortcuts (Section 5).