Memorizing is Not Enough: Deep Knowledge Injection Through Reasoning

Ruoxi Xu[1,2], **Yunjie Ji**[3,4,*], **Boxi Cao**[1], **Yaojie Lu**[1], **Hongyu Lin**[1], **Xianpei Han**[1], **Ben He**[2,*], **Yingfei Sun**[2], **Xiangang Li**[3,4], **Le Sun**[1,*]
[1]Chinese Information Processing Laboratory, Institute of Software, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences [3]Beike Inc. [4]a-m-team
{ruoxi2021, boxi2020, luyaojie, hongyu, xianpei, sunle}@iscas.ac.cn
{jiyunjie001, lixiangang002}@ke.com
{benhe, yfsun}@ucas.edu.cn

Abstract

Although large language models (LLMs) excel in knowledge recall and reasoning, their static nature leads to outdated information as the real world evolves or when adapting to domain-specific knowledge, highlighting the need for effective knowledge injection. However, current research on knowledge injection remains superficial, mainly focusing on knowledge recall and extraction. This paper proposes a four-tier knowledge injection framework that systematically defines the levels of knowledge injection: recall, extraction, reasoning, and association. Based on this framework, we introduce DeepKnowledge, a synthetic experimental testbed designed for fine-grained evaluation of the depth of knowledge injection across three knowledge types (novel, incremental, and updated). We then explore various knowledge injection scenarios and evaluate the depth of knowledge injection for each scenario on the benchmark. Experimental results reveal key factors to reach each level of knowledge injection for LLMs and establish a mapping between the levels of knowledge injection and the corresponding suitable injection methods, aiming to provide a comprehensive approach for efficient knowledge injection across various levels. The code is available at https://github.com/icip-cas/Knowledge-Learning-Toolkits.

1 Introduction

LLMs have the remarkable ability to capture vast amounts of factual knowledge from extensive pre-training data (AlKhamissi et al., 2022; Cao et al., 2021; Meyer et al., 2023). However, their static nature leads to knowledge becoming outdated as real-world information evolves or when adapting to new and private domain knowledge (Wang et al., 2024b). To mitigate these issues, continual pre-training on updated or domain-specific documents has become a common strategy (Zhang et al., 2023; Jang et al., 2022), aiming to refresh LLMs' knowledge and tailor it to specific areas of expertise.

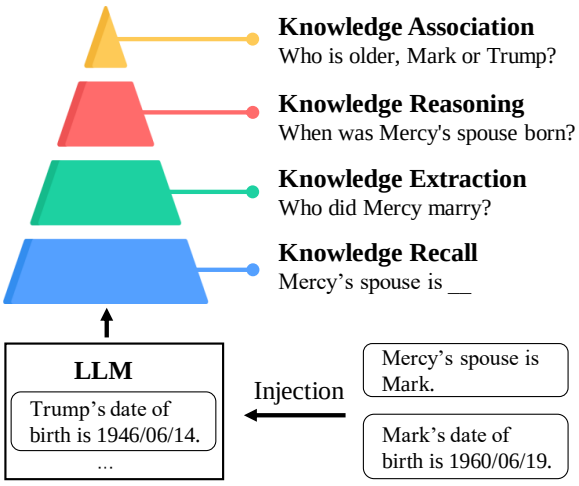

Figure 1: An illustration of the four-layer knowledge injection framework. This hierarchical framework provides a finer-grained approach to injecting knowledge into LLMs, ranging from basic recall to joint reasoning between new knowledge and pre-existing knowledge.

Knowledge injection progresses as a continuum, not a binary transition (Hu et al., 2023). However, current research on remains superficial, mainly focusing on knowledge recall and extraction. For instance, Carlini et al. (2021); Cao et al. (2024) assess knowledge recall through text completion tasks and Chang et al. (2024); Allen-Zhu and Li examine extraction via rephrased questions. Superficial knowledge struggles to support reasoning tasks, leading to under-performance in scenarios that require deep reasoning (Allen-Zhu and Li, 2023). Therefore, it is critical to conduct a systematic investigation about knowledge injection levels and establish a mapping between injection levels and suitable knowledge injection methods.

To this end, this paper proposes a four-layer knowledge injection framework, systematically defining the four key levels of knowledge injection. As illustrated in Figure 1, we divide the knowledge

*Corresponding author.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 28682–28693

injection process into four levels: 1) **Knowledge recall**: The model's ability to recall and restate the injected knowledge in its original form. 2) **Knowledge extraction**: The model's ability to correctly extract knowledge under various expressions. 3) **Knowledge Reasoning**: The model's ability to apply the injected knowledge in reasoning tasks. 4) **Knowledge Association**: The model's ability to jointly apply the injected knowledge and pre-existing knowledge in reasoning tasks.

To investigate the interactions between these layers, we develop a synthetic experimental testbed called DeepKnowledge. DeepKnowledge offers a four-tier evaluation of knowledge injection effectiveness, based on four distinct knowledge types. Specifically, the evaluation aligns with the knowledge injection levels in Figure 1: recall (Level 1), extraction (Level 2), 1-3 step reasoning (Level 3) and association (Level 4). This hierarchical evaluation allows for a more nuanced understanding of the challenges involved in integrating knowledge into LLMs, from basic recall to complex multi-step reasoning. Besides, DeepKnowledge incorporates four knowledge types: pre-existing, novel, incremental and updated knowledge. Knowledge is unique, and non-recursive to ensure valid multi-step reasoning.

Based on this setup, we systematically investigate the boundaries of knowledge injection under various knowledge injection scenarios. Our findings reveal key factors influencing the achievement of each level of knowledge injection in LLMs: 1) Repetitive learning enables rapid recall of isolated knowledge; 2) Knowledge diversity is critical for transitioning from mere recall to retrievable knowledge representation; 3) Explicit reasoning patterns link isolated knowledge for reasoning and enable generalization to new entities and deep reasoning; 4) LLMs excel at shallow knowledge association, but require explicit reasoning to forge deep connections. We infer that new knowledge must be interconnected through reasoning mechanisms to facilitate generalized knowledge reasoning.

Furthermore, we conducted ablation experiments to identify the key factors affecting the efficiency of knowledge injection. We analyzed the impact of knowledge types, data formulation and diversity on the effectiveness of knowledge injection and provided a recipe to achieve efficient knowledge injection at various levels, which could be valuable for future research.

To summarize, we make the following contributions:

- We proposed a four-layer knowledge injection framework, including knowledge recall, extraction, reasoning and association.
- We developed DeepKnowledge, a comprehensive evaluation testbed designed to assess knowledge application across different injection levels.
- We established a systematic mapping between knowledge injection levels and suitable injection methods.

2 DeepKnowledge

In this section, we present a comprehensive experimental testbed, DeepKnowledge. Aligned with the knowledge injection framework in Figure 1, DeepKnowledge enables systematic evaluation across four knowledge injection levels. The following provides a detailed description of the construction process of DeepKnowledge.

2.1 Knowledge Acquisition

Pre-existing Knowledge Filter To ensure the validity of our benchmark, we apply three filtering criteria to pre-existing factual knowledge: 1) Uniqueness: For each subject-relation pair, the object must be unique. 2) Non-recursiveness: Facts should not be recursive, i.e., the subject and object cannot be identical. 3) Multi-step Reasoning: The knowledge must be suitable for constructing multi-step reasoning tasks.

Specifically, we first get factual knowledge from WikiFactDiff (Khodja et al., 2024) and MQuAKE (Zhong et al., 2023). We then filter out facts containing special characters or empty values in the subject or object, as well as those with non-unique answers or recursive structures. This ensures that only high-quality, valid facts remain in the dataset. Next, we analyze the distribution of fact chains and manually select 16 relationship groups that are critical for reasoning tasks, as illustrated in Appendix Figure 7. Finally, we enable the model to recall facts in a 3-shot setting, retaining only the facts that the model provides correct answers as pre-existing knowledge. In total, we curate a set of 26,477 valid and reliable facts.

Synthetic Knowledge Generation To ensure that the injected knowledge contains novel information that was not seen during the pretraining phase,

we construct a synthetic knowledge dataset. To eliminate confounding factors, the synthetic knowledge is designed to match the distribution of factual knowledge. Specifically, based on the relationship types selected from the factual knowledge, we first generate fictional entity names using LLMs. For example, when generating names for entities belonging to the category of "location," we combine a random word (e.g., "Frank") with a geographic term (e.g., "town") to create an entity name such as "FrankTown." These fictional entities are then assigned the same relationships as those in the factual knowledge to form synthetic facts. As a result, we generate a set of 109860 synthetic facts.

2.2 Test Cases Generation

Recall Test Cases Generation We define the shallowest level of knowledge injection as the model's ability to retain and recognize the original content of the knowledge during training. Specifically, we focus on sentences from the training corpus that involve specific knowledge. For these sentences, the object is removed, transforming the sentence into a cloze-style question. The task for the model is to predict the correct object, which serves as the answer to the cloze question. This method effectively measures the model's ability to recall and reconstruct knowledge based on its training data, providing a baseline for evaluating its memory and knowledge retention capabilities.

Extraction Test Cases Generation We define knowledge extraction as the model's ability to extract knowledge from various semantically equivalent formulations. Specifically, we take the cloze questions from recall-level test cases and rephrase them into ten semantically equivalent questions using LLMs. These rephrased questions are then used as the input, with the original answer retained as the correct response. This approach assesses the model's capacity to recognize and extract knowledge across different expressions while ensuring that the underlying semantic content remains consistent.

Reasoning Test Cases Generation We define reasoning as the induction and application of inference rules. Specifically, we first define two fundamental reasoning rules: combination, which involves multi-hop knowledge aggregation, and comparison, which focuses on comparing the magnitude of knowledge. Each reasoning rule is considered a single step of reasoning. To construct

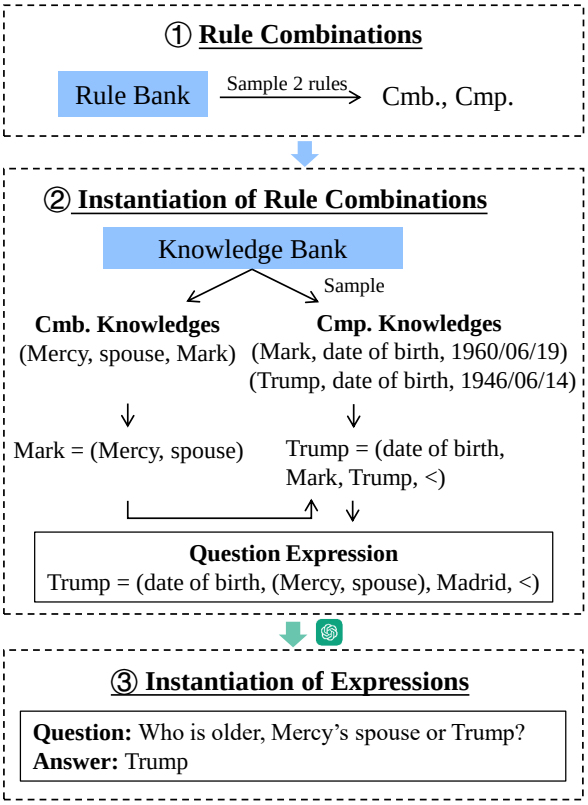

Figure 2: Construction pipeline for reason test cases, which involves three steps: 1) multi-step rule combinations, 2) instantiation of rule combinations to knowledge expressions following the rule definition, 3) instantiation of knowledge expressions to questions using LLMs.

an n-step reasoning problem, we sample n reasoning rules as the foundation of the problem, and randomly sample knowledge to populate each reasoning step to form an expression. Finally, the expression is translated into a fluent natural language question using GPT-4. The construction pipeline for reason test cases is illustrated in Figure 2. The detailed rule definitions and prompts can be found in the Appendix A.

Association Test Cases Generation We define association as the reasoning process that connects newly injected knowledge with pre-existing knowledge in LLMs. Specifically, we adopt an approach similar to question generation for reasoning tasks, with the key distinction that the constructed questions must incorporate both new and existing knowledge.

3 Experiment Settings

3.1 Knowledge Types

The correlation between newly injected knowledge and the pre-existing knowledge in LLMs may influence the effectiveness of the injection. There-

fore, we systematically explore three knowledge paradigms that encompass possible types of new knowledge (Khodja et al., 2024): 1) Novel Knowledge, which introduces entirely new information about emerging entities (e.g., a newly proposed scientific theory); 2) Incremental Knowledge, which expanding existing entities with supplemental facts (e.g., new publications by established authors); and 3) Updated Knowledge, where outdated facts are replaced with current information (e.g., a sports team appointing a new coach).

3.2 Injection Scenarios

We explored five distinct knowledge injection scenarios: 1) Duplicate: Repeating the same knowledge multiple times without modification. 2) Vanilla Paraphrase: Rephrasing the injected knowledge using a LLM, altering its expression. 3) Style-enhanced Paraphrase: Rephrasing the injected knowledge with style variations, where the rephrasing style is randomly selected from a predefined style bank (details provided in Appendix B). 4) Single-step Implicit Reasoning: Combining the rephrased knowledge with a single-step reasoning question and its corresponding answer. 5) Single-step Explicit Reasoning: Combining the rephrased knowledge with a single-step reasoning question, a detailed reasoning process, and the corresponding answer. For all these knowledge injection scenarios, we ensured that each knowledge was injected 20 times, regardless of its form, to eliminate any potential influence of training data size.

3.3 Test Settings

We evaluate the model's performance under three distinct in-context learning settings: 1) 0-shot, where LLMs generate answers without any prior context or examples. 2) 3-shot, where three relevant examples are selected based on various criteria: for recall, examples share the same relation type as the test question; for extraction, examples contain answers related to the same entity type as the test question; and for reasoning and association, examples exhibit similar reasoning structures to the test question. 3) 3-shot CoT: which mirrors the 3-shot setting, but includes examples that explicitly demonstrate reasoning processes.

3.4 Evaluation Metrics

To evaluate the model's performance in answering free-text questions, we leverage tailored regular expression-based parsing techniques. Specifically, for recall and extraction questions, the model's response is deemed correct if the ground truth is contained within the first sentence of its generated text. For reasoning and association questions, we employ a prompt beginning with "Answer:" to guide the model in providing explicit options, extracting the initials of its generated answers. If these align fully with the ground truth, the response is considered accurate.

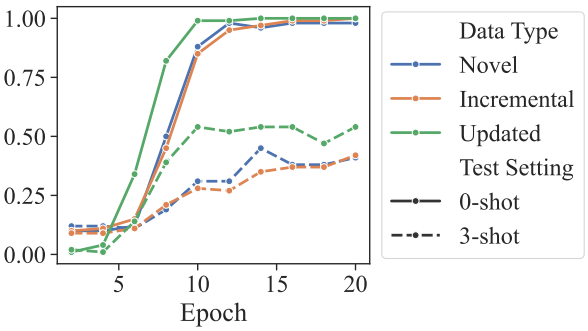

Figure 3: The knowledge recall score during the Duplicate Injection process. Repetitive learning enables knowledge recall.

3.5 Training Details

We adopted continued pretraining (CPT) to inject new knowledge into LLMs, motivated by existing research that suggests supervised fine-tuning may induce hallucinations (Gekhman et al., 2024). Our primary experiments were conducted on the LLaMA 3-8B model. To ensure stability and robust generalization, we trained LLMs using a balanced mixture of training data and general instructions at a 1:1 ratio, with a learning rate of 3e-5.

4 What's the key to reach each level of knowledge injection for LLMs?

In this section, we explore various knowledge injection scenarios and evaluate their knowledge injection levels on the DeepKnowledge testbed. Our experiments reveal that different knowledge injection strategies are required to attain various levels of knowledge injection in LLMs. In the following, we will illustrate the experiment findings to reach the above conclusion.

4.1 Knowledge Recall

Repetitive learning enables LLMs to memorize context-dependent and isolated knowledge.

Injection Scenario	Test	Reason (2 steps)			Reason (3 steps)		
		0S	3S	3S-CoT	0S	3S	3S-CoT
-	Old	35.7	37.0	64.0	26.3	31.0	55.3
Duplicate	Novel	11.3	24.7	3.3	8.3	26.7	3.7
	Updated	9.0	18.3	4.3	7.3	19.7	3.3
Style-enhanced Paraphrase	Novel	22.3	25.7	31.3	28.3	29.3	24.7
	Updated	24.0	26.3	30.3	25.3	20.0	31.0
Single-step Implicit Reason	Novel	28.7	39.0	34.3	41.7	34.0	31.7
	Updated	24.7	26.7	35.3	38.3	40.0	33.7
Single-step Explicit Reason	Novel	21.3	21.3	41.0	20.7	24.0	49.3
	Updated	33.7	30.3	52.0	24.7	30.0	57.3

Table 1: The knowledge reasoning score of LLMs across various training and test settings. "0S", "3S" and "3S-CoT" correspond to the 0-shot, 3-shot, and 3-shot CoT test settings, respectively. Scores are presented on a green-white-red scale. From the table, we can observe that newly injected knowledge is often isolated and needs to be connected through explicit reasoning, enabling generalization to new entities and deep reasoning.

From Figure 3, we observe that under the duplicate injection scenario, the knowledge recall scores in the 0-shot setting steadily increase with the number of repetitions, stabilizing at approximately 95 across different knowledge types. This suggests that LLMs are highly effective at memorizing training data when specific knowledge is frequently repeated during training. Consequently, for knowledge that requires simple recall in its original form, repeated exposure during training is sufficient to support the recall process.

However, we also note the following, as shown in Figure 3 and Table 1: 1) Under the duplicate injection scenario, the knowledge recall score in the 3-shot setting is significantly lower than in the 0-shot setting. 2) The duplicate injection results in very low scores for knowledge extraction and reasoning. This indicates that, at this stage, while the model can recall knowledge almost perfectly, the memorized knowledge remains unstable, easily influenced by context, and lacks connections to other knowledge.

4.2 Knowledge Extraction

Diverse and heterogeneous expressions bridge knowledge recall and extraction.

From Figure 4, we observe the following: 1) Under duplicate injection scenarios, LLMs' knowledge extraction score consistently remains around 20 across all knowledge types, even though document perplexity is minimized; 2) LLMs' knowledge extraction score improves significantly when presenting knowledge through diversified linguistic expressions. 3) With the same data augmentation factor, the style-enhanced paraphrasing method achieves substantially better knowledge extraction improvement compared with standard LLM-based paraphrasing approaches. This suggests that diverse knowledge expressions are essential for enabling effective knowledge extraction during the injection process. Knowledge expression divergence serves as the critical enabler for extraction capability enhancement. Therefore, to enable knowledge extraction, it is essential to present knowledge in diverse and varied expressions. This diversity is key to improving the model's extraction capability.

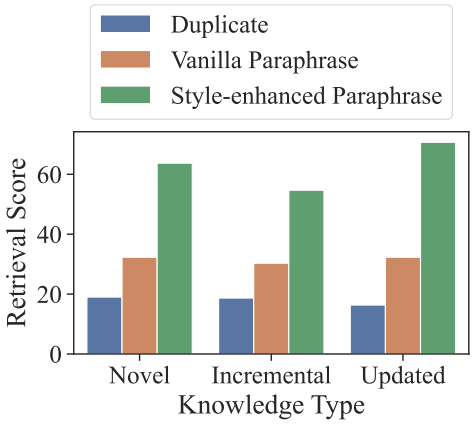

Figure 4: The knowledge extraction score of LLMs under different knowledge injection scenarios. Diverse and divergent expressions are key to effective knowledge extraction.

Training	Test	Association (2 steps)			Association (3 steps)		
		0S	3S	3S-CoT	0S	3S	3S-CoT
Duplicate	Novel&Old	15.7	27.3	8.3	12.3	14.3	5.0
	Inc.&Old	13.0	28.0	2.7	10.7	24.0	4.7
	Updated&Old	15.7	35.3	7.7	9.7	31.0	6.0
Style-enhanced parapharse	Novel&Old	28.7	27.7	49.7	32.7	26.7	27.0
	Inc.&Old	31.0	29.7	46.3	26.3	28.3	28.3
	Updated&Old	28.0	32.3	41.0	19.7	27.7	33.3
Single-step Implicit Reason	Novel&Old	27.3	35.0	42.0	32.0	37.0	31.7
	Inc.&Old	30.3	43.3	41.7	25.0	33.7	30.0
	Updated&Old	27.7	30.3	39.7	29.3	31.7	35.3
Single-step Explicit Reason	Novel&Old	26.7	40.0	48.3	17.7	26.0	38.0
	Inc.&Old	24.0	32.7	40.7	27.0	29.0	39.0
	Updated&Old	30.3	32.7	48.3	29.0	40.0	57.3

Table 2: The knowledge association score of LLMs across different knowledge types. "Inc." refers to incremental knowledge. "0S", "3S" and "3S-CoT" correspond to the 0-shot, 3-shot, and 3-shot CoT test settings, respectively. From the table, we can see that LLMs excel at shallow knowledge association, but require explicit reasoning to forge deep connections.

4.3 Knowledge Reasoning

Explicit reasoning patterns link isolated knowledge for reasoning and enable generalization to new entities and deep reasoning.

As demonstrated in Sections 4.1 and 4.2, newly injected knowledge is often isolated, making it challenging for LLMs to perform multi-knowledge reasoning. To enhance the model's reasoning capability, we incorporated both single-step implicit and explicit reasoning data into the training set. We then tested the model's ability to generalize reasoning capabilities to novel knowledge and multi-step reasoning scenarios.

The experimental results presented in Table 1 reveal four key observations: 1) Implicit reasoning patterns enhance zero-shot performance for multi-step reasoning tasks; 2) Explicit reasoning patterns improve multi-step reasoning performance in the 3-shot CoT setting; 3) Explicit reasoning data achieves a higher upper-bound performance in multi-step reasoning compared to implicit reasoning data (57.3 vs 41.7); 4) Single-step explicit training demonstrates effective generalization to novel entities and multi-step reasoning scenarios. Thus, connecting new knowledge through a reasoning mechanism is crucial for facilitating knowledge reasoning. This connection enables immediate application and enhances generalization across various contexts.

4.4 Knowledge Association

LLMs excel at shallow knowledge association, but require explicit reasoning to forge deep connections.

As shown in Table 2 for shallow knowledge associations (2-step), the paraphrase injection yields knowledge association scores around 45, whereas duplicate injection achieves scores near 5, demonstrating significant improvement. This suggests that rigid, highly redundant injection patterns damage shallow-level reasoning capabilities, while diversified knowledge injection better preserves existing reasoning capacity and facilitates generalization to joint reasoning across old and new knowledge.

For deep knowledge associations (3-step), explicit reasoning pattern injection restores knowledge integration scores to levels comparable to the base model without knowledge injection. This evidence indicates that simple recall mechanisms struggle to synthesize new and existing knowledge. By incorporating explicit reasoning patterns during injection, LLMs can develop systematic knowledge association capabilities.

Train	Ratio	Recall		Extraction		Reason (1 step)			Reason (2 steps)			Reason (3 steps)		
		0S	3S	0S	3S	0S	3S	3S-CoT	0S	3S	3S-CoT	0S	3S	3S-CoT
Duplicate	2:1	99.0	32.7	8.7	22.0	6.0	1.3	1.3	4.3	2.0	1.3	1.3	0.0	0.7
	1:1	97.7	40.3	19.0	23.3	24.0	30.0	5.0	11.3	25.3	3.0	8.3	27.3	4.3
	1:2	98.7	33.3	11.7	28.3	13.0	20.7	18.0	11.3	24.7	17.0	7.7	15.7	10.3
Style-enhanced parapharse	2:1	92.7	61.0	58.3	54.3	33.3	26.7	36.7	17.3	27.7	34.0	21.3	19.3	34.7
	1:1	93.7	56.0	63.7	48.7	27.3	33.7	43.3	22.0	26.3	31.3	28.3	29.0	25.3
	1:2	93.7	60.7	75.3	54.7	25.0	22.0	46.7	18.0	24.7	29.7	19.0	19.7	40.7
Implicit Reason	2:1	89.7	66.3	58.7	48.7	53.0	62.7	28.7	35.3	32.0	13.3	39.0	40.7	24.3
	1:1	92.0	59.0	65.7	50.3	37.7	55.3	53.0	29.3	39.3	34.3	41.7	33.7	32.3
	1:2	97.0	69.0	74.3	61.0	60.3	58.0	57.0	34.7	33.0	32.3	46.0	41.3	45.3
Explicit Reason	2:1	93.3	86.3	86.7	76.7	51.0	43.3	79.0	32.7	34.3	0.0	27.3	26.7	6.3
	1:1	97.3	73.7	74.0	67.0	44.7	37.3	64.0	21.0	21.0	41.0	21.0	24.0	49.3
	1:2	95.7	79.3	85.0	75.3	40.7	36.7	64.3	29.0	33.7	54.3	26.0	33.0	54.7

Table 3: The novel knowledge injection score of LLMs under different ratios of general instructions. As shown in the table, an adequate amount of general instructions is crucial for knowledge reasoning.

5 Error Analysis

To gain a deeper understanding of the challenges and bottlenecks of knowledge injection, we conduct a taxonomy analysis of model failures observed under explicit reasoning injection scenarios. This approach helps us identify the key issues that need to be addressed. From Figure 5, we can observe the following findings:

1) For novel knowledge, wrong reason paths are the primary factor leading to incorrect answers in complex reasoning tasks. Empirical evidence indicates that over 50% of errors stem from incorrect problem decomposition. This observation aligns with the intuitive understanding that the core challenge in complex reasoning tasks lies in properly grasping the structure of the problem and effectively breaking it down into manageable sub-tasks. When this decomposition is flawed, it becomes difficult for the model to proceed with the reasoning in a systematic and coherent manner.

2) For updated knowledge, wrong knowledge is one of the leading causes of errors in complex reasoning. Notably, compared to novel knowledge, errors arising from faulty recall of updated knowledge are more prevalent. This is because introducing new information that contradicts or differs from the model's existing knowledge often leads to hallucinations. The model struggles to apply the new knowledge consistently across various reasoning levels, resulting in inconsistencies.

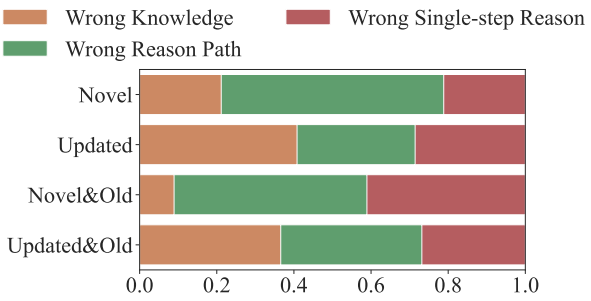

Figure 5: Proportions of error causes of complex reason tasks.

6 Ablation Study

In this section, we conduct ablation experiments to identify the key factors influencing knowledge injection efficiency. We analyze the impact of knowledge type, data formulation, and diversity on the effectiveness of knowledge injection, aiming to provide a recipe for efficient knowledge injection across different levels.

6.1 Effect of Injected Knowledge Types

We investigate how knowledge types affect injection efficiency, focusing on three types: novel, incremental and updated. While all introduce new information, they differ in the model's prior familiarity with involved entities. Our key findings are: 1) LLMs exhibit comparable recall capabilities across knowledge types (>95% scores after repeated training), as shown in Figure 3. This observation aligns with our expectation, as knowledge recall is a relatively simple task for LLMs. They

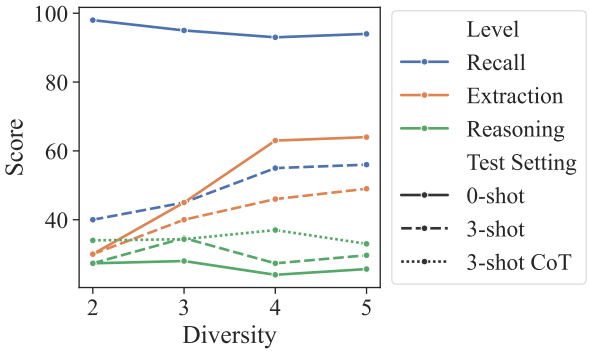

Figure 6: Effect of the diveristy of injected knowledge.

reliably memorize new textual knowledge regardless of its relation to existing information. 2) Updated knowledge significantly outperforms novel knowledge in complex tasks, including extraction, reasoning and association, as shown in Figure 4 and Table 1, 2. We attribute this to the model's pre-existing reasoning frameworks for updated entities, consistent with Wang et al. (2024a)'s findings. Novel knowledge requires additional generalization as its entities lack established associations.

6.2 Effect of Data Formulation

We investigate mixing ratios between knowledge and general instructions (1:2, 1:1, 2:1) using new knowledge exemplars (Cheng et al., 2023). Our experiments results in Table 3 demonstrate that general data significantly enhances knowledge application, particularly in complex reasoning. The results reveal a positive correlation between general data proportion and success rates in multi-step reasoning tasks. We therefore implement a balanced 1:1 ratio for optimal efficiency and effectiveness in knowledge injection.

6.3 Effect of Diversity of Injected Knowledge

We explored the impact of knowledge diversity on knowledge application by gradually increased the diversity of textual representations of the same knowledge (2-5 variants) with fixed 20 training iterations. We used paraphrased new knowledge injection as an example. The experimental results are shown in Figure 6. Our findings indicate that, greater diversity in phrasing is positively correlated with the model's ability to retrieve knowledge up to 4 variants. However, beyond this range, further increasing phrasing diversity does not improve the model's knowledge application ability. This identifies an optimal diversity threshold for effective knowledge application, providing practical guidance for knowledge injection strategies in LLMs.

7 Related Work

Knowledge injection is not a binary distinction, but rather a process of gradual transition from 0 to 1 (Yang et al., 2021; Wang et al., 2024b). Existing work on knowledge injection typically focuses on shallow-level tasks like recall and extraction, with limited exploration of complex reasoning scenarios. Prior work evaluates recall through text completion (Carlini et al., 2021; Cao et al., 2024; Chang et al., 2024) and extraction via question answering (Jiang et al., 2024; Allen-Zhu and Li; Ovadia et al., 2023; Zhu et al., 2024), while recent studies explore basic single-step operations like comparison and combination (Lu et al., 2024; Allen-Zhu and Li, 2023; Wang et al., 2024a). However, the generalization boundaries of knowledge injection in complex and joint reasoning tasks remain unclear. To address this, we propose a four-level injection framework and develop a systematic benchmark for evaluating method limitations.

Furthermore, current knowledge injection work tends to focus on single types of knowledge while overlooking the influence of the relationship between injected knowledge and the model's pre-existing knowledge on injection efficiency. For instance, Chang et al. (2024); Allen-Zhu and Li, 2023) inject entirely novel knowledge, whereas other studies focus on updated knowledge (Wang et al., 2024b; Zhang et al., 2024). To bridge this gap, we formally define three fundamental knowledge types: novel, incremental, and updated, enabling systematic analysis of how knowledge relationships affect injection efficiency.

8 Conclusion

In this paper, we propose a four-layer knowledge injection framework, which includes knowledge recall, extraction, reasoning, and association. This framework is designed to enable a granular investigation into the depth of newly injected knowledge in LLMs. Building upon this framework, we further construct DeepKnowledge, a multi-level evaluation benchmark designed to systematically assess knowledge injection methods across distinct cognitive layers. Using this benchmark, we evaluate the effectiveness of various knowledge injection methods. The experimental results highlight that achieving effective knowledge injection across different levels requires careful attention to different core training factors, providing valuable guidance for efficient knowledge injection.

Limitation

Due to resource and time constraints, our experiments were limited to the LLAMA3-8B model. We focused on continuous pre-training as the primary method for knowledge injection, as it is a well-established approach that helps mitigate hallucinations. Additionally, our study only explored reasoning operations involving the comparison and combination of atomic operations. Future work will expand the research to include a broader range of model sizes, knowledge injection methods, and reasoning paradigms.

Acknowledgment

We sincerely thank the reviewers for their insightful comments and valuable suggestions. This work was supported by Beijing Natural Science Foundation (L243006), Beijing Municipal Science and Technology Project (Nos. Z231100010323002), the Natural Science Foundation of China (No. 62306303, 62476265, 62272439) and the Basic Research Program of ISCAS (Grant No. ISCAS-JCZD-202303).

References

Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona Diab, and Marjan Ghazvininejad. 2022. A review on language models as knowledge bases. *arXiv preprint arXiv:2204.06031*.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.1, knowledge storage and extraction. In *Forty-first International Conference on Machine Learning*.

Zeyuan Allen-Zhu and Yuanzhi Li. 2023. Physics of language models: Part 3.2, knowledge manipulation. *arXiv preprint arXiv:2309.14402*.

Boxi Cao, Hongyu Lin, Xianpei Han, Le Sun, Lingyong Yan, Meng Liao, Tong Xue, and Jin Xu. 2021. Knowledgeable or educated guess? revisiting language models as knowledge bases. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1860–1874.

Boxi Cao, Qiaoyu Tang, Hongyu Lin, Shanshan Jiang, Bin Dong, Xianpei Han, Jiawei Chen, Tianshu Wang, and Le Sun. 2024. Retentive or forgetful? diving into the knowledge memorizing mechanism of language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14016–14036.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

Hoyeon Chang, Jinho Park, Seonghyeon Ye, Sohee Yang, Youngkyung Seo, Du-Seong Chang, and Minjoon Seo. 2024. How do large language models acquire factual knowledge during pretraining? *arXiv preprint arXiv:2406.11813*.

Daixuan Cheng, Shaohan Huang, and Furu Wei. 2023. Adapting large language models via reading comprehension. In *The Twelfth International Conference on Learning Representations*.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. 2024. Does fine-tuning llms on new knowledge encourage hallucinations? *arXiv preprint arXiv:2405.05904*.

Linmei Hu, Zeyi Liu, Ziwang Zhao, Lei Hou, Liqiang Nie, and Juanzi Li. 2023. A survey of knowledge enhanced pre-trained language models. *IEEE Transactions on Knowledge and Data Engineering*.

Joel Jang, Seonghyeon Ye, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, Jungkyu Choi, and Minjoon Seo. 2022. Towards continual knowledge learning of language models. In *10th International Conference on Learning Representations, ICLR 2022*. International Conference on Learning Representations.

Zhengbao Jiang, Zhiqing Sun, Weijia Shi, Pedro Rodriguez, Chunting Zhou, Graham Neubig, Xi Victoria Lin, Wen-tau Yih, and Srinivasan Iyer. 2024. Instruction-tuned language models are better knowledge learners. *arXiv preprint arXiv:2402.12847*.

Hichem Ammar Khodja, Frédéric Bechet, Quentin Brabant, Alexis Nasr, and Gwénolé Lecorvé. 2024. Wikifactdiff: A large, realistic, and temporally adaptable dataset for atomic factual knowledge update in causal language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17614–17624.

Xingyu Lu, Xiaonan Li, Qinyuan Cheng, Kai Ding, Xuan-Jing Huang, and Xipeng Qiu. 2024. Scaling laws for fact memorization of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11263–11282.

Lars-Peter Meyer, Claus Stadler, Johannes Frey, Norman Radtke, Kurt Junghanns, Roy Meissner, Gordian Dziwis, Kirill Bulert, and Michael Martin. 2023. Llm-assisted knowledge graph engineering: Experiments with chatgpt. In *Working conference on Artificial Intelligence Development for a Resilient and Sustainable Tomorrow*, pages 103–115. Springer Fachmedien Wiesbaden Wiesbaden.

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. 2023. Fine-tuning or retrieval? comparing knowledge injection in llms. *arXiv preprint arXiv:2312.05934*.

Boshi Wang, Xiang Yue, Yu Su, and Huan Sun. 2024a. Grokked transformers are implicit reasoners: A mechanistic journey to the edge of generalization. *arXiv preprint arXiv:2405.15071*.

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. 2024b. Knowledge editing for large language models: A survey. *ACM Computing Surveys*, 57(3):1–37.

Jian Yang, Xinyu Hu, Gang Xiao, and Yulong Shen. 2021. A survey of knowledge enhanced pre-trained models. *arXiv preprint arXiv:2110.00269*.

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, et al. 2024. A comprehensive study of knowledge editing for large language models. *arXiv preprint arXiv:2401.01286*.

Zihan Zhang, Meng Fang, Ling Chen, Mohammad-Reza Namazi-Rad, and Jun Wang. 2023. How do large language models capture the ever-changing world knowledge? a review of recent advances. *arXiv preprint arXiv:2310.07343*.

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*, pages 15686–15702.

Tongyao Zhu, Qian Liu, Liang Pang, Zhengbao Jiang, Min-Yen Kan, and Min Lin. 2024. Beyond memorization: The challenge of random memory access in language models. *arXiv preprint arXiv:2403.07805*.

A DeepKnowledge

A.1 Examples

In Table 4, we present examples of DeepKnowledge.

A.2 Inference Rules

We conceptualize reasoning as the induction and application of inference rules. Specifically, we adopt the following inference rules:

- Combination: The two-hop combination rule is defined as follows:

$$\begin{aligned} &\forall h, b, t \in E, \forall r_1, r_2 \in R, \\ &(h, r_1, b) \wedge (b, r_2, t) \Rightarrow t = (h, r_1, r_2) \end{aligned} \tag{1}$$

  Here, h, b, t represent entities, and r_1, r_2 are relations.

- Comparison: The comparison rules are as follows:

$$\begin{aligned} \forall e_1, e_2 \in E, \quad \forall a \in A, \quad \forall v_1, v_2 \in V, \\ (e_1, a, v_1) \wedge (e_2, a, v_2) \wedge v_1 < v_2 \\ \Rightarrow e1 = (a, e_1, e_2, a <) \\ (e_1, a, v_1) \wedge (e_2, a, v_2) \wedge v_1 > v_2 \\ \Rightarrow e2 = (a, e_1, e_2, a >) \end{aligned} \tag{2}$$

  Here, e_1, e_2 are entities, a is an attribute, and v_1, v_2 are values, with the comparison being based on the relationship between the values.

A.3 Prompts

We provide the prompts we used to generate test cases for knowledge reasoning in the following.

> Define two basic operation rules:
> Comparison: $(e1, a, v1) \wedge (e2, a, v2) \wedge v1 < v2 \Rightarrow e1 = (a, e1, e2, <), e2 = (a, e1, e2, >)$
> Combination: $(h, r1) \Rightarrow t = (h, r1)$
> Please complete the following task in the format given in the example.
> Task: Given an expression formed by basic operation rules, explain it step by step from the innermost rule to the outermost rule to create a complex reasoning question. The following requirements must be met: 1. The question must be purely a question and cannot include the reasoning results of the rules, nor can it change, add, or omit any of facts beyond the rule. 2. Only output a question in

Level	Rules	Knowledge	Question	Answer
Recall	-	(Mercy, language of work or name, English)	Mercy speaks	English
Extraction	-	(Mercy, language of work or name, English)	What is the language of work or name for Mercy?	English
Single-step Reason	Cmb.	(My Sweet Lord, performer, George Harrison), (George Harrison, country of citizenship, United Kingdom)	What's the country of citizenship of the performer of the song "My Sweet Lord"?	United Kingdom
Two-steps Reason	Cmp., Cmb.	(12th Magritte Awards, country, Belgium), (Belgium, population, 11584008), (Madrid, population, 3280782)	Which one has a smaller population, the country where the 12th Magritte Awards took place or Madrid?	Madrid
Three-steps Reason	Cmb., Cmb., Cmp.	(Kimberly Gary Sutton, spouse, John Gerald Price), (John Gerald Price, country of citizenship, Aliceville), (Aliceville, population, 150000), (Virginiaopolis, population, 8504231)	Which region has a smaller population, the country of citizenship of the spouse of Kimberly Gary Sutton or Virginiaopolis?	Aliceville

Table 4: Examples of the benchmark designed to evaluate five levels of knowledge application depth. The terms "Cmb." and "Cmp." denote "Combination" and "Comparison," respectively, which represent the fundamental reasoning operations in evaluation tasks.

only one version without any additional explanations. 3. The question should be fluent, easy to read, and concise.

Expression: [['Joe Jacob Addington', 'spouse'], 'country of citizenship']

Question: What's the country of citizenship of the spouse of Joe Jacob Addington?

Expression: ['retirement age', 'Celloria', 'Careerlandia', '<']

Question: Which one has a lower retirement age, Keithville or Katherineville?

Expression: [['inception', 'FC Lokomotiv 1929 Sofia', 'FC Rapid 1923', '<'], 'sport']

Question: What sport do the club that was established earlier between FC Lokomotiv 1929 Sofia and FC Rapid 1923 play?

Expression: ['female population', ['Leroy Christopher Austin', 'country of citizenship'], 'Edwardville', '<']

Question: Which region has a smaller female population, Leroy Christopher Austin's country of citizenship or Edwardsville?

Expression: ['male population', ['male population', ['male population', 'Brianville', 'Evasville', '>'], 'Ellenborough', '<'], 'Gregorian Chronicles', '<']

Question: Compare the male population of Brianville and Evasville, and select the region with the larger male population. Then, compare this region's male population with that of Ellenborough and select the region with the smaller population. Which one has a smaller male population, this region or Gregorian Chronicles?

Expression: {expression}

Answer: {answer}

Question:

A.4 Fact Chains

The fact chains, including entity and relationship types, are presented in Figure 7.

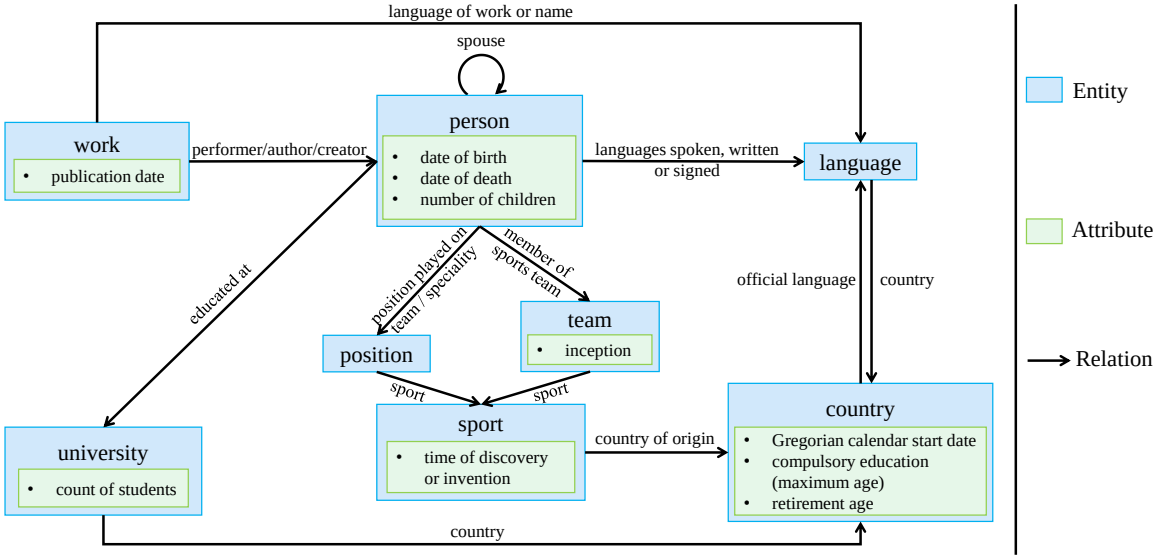

Figure 7: Fact chains of our benchmark.

B Experiment Setting

B.1 Styles

We manually defined a style bank to enhance the diversity of paraphrased text, which includes four main categories:

- Text Genre: textbook, news, academic paper, lyrics, dialogue, speech, story, summary
- Text Type: question-answer, exclamation
- Text Sentiment: positive, negative
- Text Formality: informal, formal