KNOWLEDGE AUGMENTATION: IN-CONTEXT OR IN-PARAMETER?

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative language models rely on knowledge augmentation to enhance their ability to complete a variety of tasks by incorporating relevant external information. The most common approach is in-context knowledge injection, where the relevant information is appended directly to the model's input. While straightforward and easy to implement, this approach is limited for complex reasoning tasks due to input length constraints and shallow integration of external and internal knowledge in language models. To overcome these limitations, we introduce an in-parameter knowledge injection method, which temporarily embeds external knowledge into the model's parameters. By injecting knowledge this way, the language models can access and reason over the information more flexibly, enhancing performance on tasks requiring sophisticated reasoning. We conduct deep explorations of both in-context and in-parameter knowledge injection, highlighting their respective advantages and limitations. Through comprehensive experiments across tasks of varying complexity, we demonstrate that in-parameter knowledge injection is particularly advantageous for complex tasks requiring deep reasoning. In contrast, in-context injection remains effective for simpler tasks where the answer can be directly extracted. Our findings provide practical guidance for selecting appropriate knowledge augmentation strategies based on the complexity of the task[1].

1 INTRODUCTION

Generative language models have significantly changed the field of natural language processing by demonstrating remarkable capabilities across a variety of tasks (Brown et al., 2020; Chowdhery et al., 2022; Touvron et al., 2023; Scao et al., 2022; Zhang et al., 2022). A critical factor contributing to their success is their ability to utilize external knowledge effectively and efficiently, thereby improving their performance on specific tasks (Lewis et al., 2020). Currently, the predominant method for knowledge augmentation is the in-context knowledge injection paradigm, wherein relevant information is appended directly to the model's input (Dong et al., 2022; Lewis et al., 2020; Levy et al., 2024). This paradigm is widely adopted due to its simplicity in implementation and immediate applicability, as it requires no modifications to the model's parameters.

However, despite its straightforward implementation, the in-context method faces limitations. Firstly, The finite length of the input context restricts the amount of external knowledge that can be incorporated (Levy et al., 2024). More critically, as shown by previous studies, language models process knowledge in input prompts and model parameters differently(Nanda et al., 2023), which means that simply adding information to the input may not be enough to activate the full power of language models' knowledge reasoning abilities. This shortcoming becomes evident in tasks that require multi-hop inference or advanced reasoning over the injected knowledge (Li et al., 2024a; Levy et al., 2024).

In contrast, an alternative yet underexplored approach is in-parameter knowledge injection(Meng et al., 2022; Li et al., 2024b). Specifically, we can temporarily inject external knowledge into language models through localized fine-tuning and restoring the original language model parameters once the task is finished. Unlike the input context, a model's parameters are not constrained by length, offering the potential to incorporate more extensive and nuanced information. By internal-

---

[1] We have open-sourced all the code, data, and models in the following anonymous GitHub link: https://anonymous.4open.science/r/In-parameter-Knowledge-Injection/

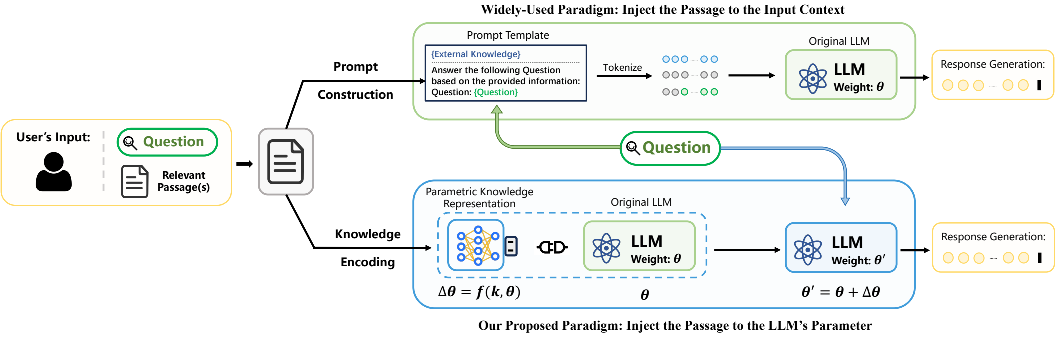

Widely-Used Paradigm: Inject the Passage to the Input Context
User's Input:
Question
Relevant Passage(s)
Prompt Construction
Prompt Template
{External Knowledge}
Answer the following Question based on the provided information:
Question: {Question}
Tokenize
Original LLM
LLM
Weight: θ
Response Generation:
Question
Knowledge Encoding
Parametric Knowledge Representation
Original LLM
LLM
Weight: θ
LLM
Weight: θ'
Response Generation:
$\Delta\theta = f(k, \theta)$
θ
$\theta' = \theta + \Delta\theta$
Our Proposed Paradigm: Inject the Passage to the LLM's Parameter

Figure 1: An illustration of in-context and in-parameter knowledge augmentation paradigms. Function f updates parameters θ by incorporating external knowledge $\mathcal{K}$, resulting in a parameter shift $\Delta\theta$ and new parameters $\theta' = \theta + \Delta\theta$.

izing external knowledge within the parameters, the language model can flexibly utilize this information and engage in deeper reasoning. Despite its potential, in-parameter knowledge injection has not been thoroughly explored. Unlike the in-context paradigm which directly places knowledge tokens into the language model's context, the in-parameter paradigm is more complex and requires updating the model's parameters. This necessitates an effective representation learning method incorporating passage-specific knowledge into the parameters while allowing the language model to apply this knowledge flexibly.

In this paper, we address this research gap by investigating circumstances under which methods, i.e., in-context or in-parameter, could produce better effectiveness in knowledge augmentation. Given that no prior studies have explored the **temporary injection** of relevant knowledge through in-parameter approaches, we propose a novel In-Parameter Knowledge Injection method. This simple yet effective method efficiently integrates external knowledge into the model's internal parameters, enhancing its reasoning capabilities regarding the injected knowledge and integrating it with other internal parametric knowledge. By generating knowledge-specific parametric representations tailored to the knowledge being injected, our proposed approach enables the model to internalize and apply the injected external information during response generation.

We design a series of progressive experiments tailored to incrementally increase both the difficulty level of the tasks and the requisite amount of reasoning needed for their resolution. This structured approach allowed us to evaluate the effectiveness of different knowledge injection methods under varying complexities. Our findings reveal a distinct trend: as task complexity and the depth of reasoning over injected knowledge rise, the in-parameter knowledge injection approach becomes increasingly advantageous. In scenarios demanding sophisticated reasoning, such as multi-document reading comprehension and multi-hop inference across multiple documents, injecting knowledge directly into the model's parameters markedly enhances performance. Conversely, the in-context approach demonstrates superior efficacy for straightforward questions or direct answer extraction from the context. This nuanced understanding of knowledge utilization underscores the critical importance of aligning the knowledge injection methodology with the specific demands of the task at hand.

In conclusion, our work provides a comprehensive analysis of the two paradigms of knowledge augmentation in language models. We highlight the strengths and limitations of each method and offer practical guidance on selecting the appropriate approach based on the task requirements. The contributions of this paper are as follows:

- We propose a simple yet effective method for in-parameter knowledge injection that leverages representation learning to incorporate external knowledge into the parameters of language models.

- We conduct extensive experiments across various scenarios to compare the performance of in-context and in-parameter knowledge injection methods. Our findings highlight the strengths and limitations of each approach, providing practical guidance on selecting the most suitable method based on task requirements.

2 PROBLEM FORMULATION OF KNOWLEDGE AUGMENTATION

Knowledge augmentation for language models involves temporarily integrating relevant external information to aid the model in executing specific tasks. This relevant knowledge is utilized solely for the duration of the task and is subsequently discarded. The main goal of knowledge augmentation is to enhance the model's performance by incorporating relevant information that is not adequately covered in its pre-training data. Two primary paradigms for injecting this knowledge into language models are in-context and in-parameter, as illustrated in Figure 1. In this section, we provide a formal definition of these methods, explaining their distinct mechanisms for injecting knowledge.

2.1 IN-CONTEXT KNOWLEDGE INJECTION PARADIGM

In the in-context knowledge injection paradigm, the relevant external knowledge is appended directly to the model's input as part of the prompt. This paradigm uses a prompt template with inserted knowledge, guiding the model to generate a response based on the provided information. Different performances in knowledge augmentation can be realized by adjusting the prompt design. For a given task, we have the query q and the external knowledge base $\mathcal{K}$. The input to the model becomes $(q, \mathcal{K})$, and the model is instructed to generate the answer a based on q and $\mathcal{K}$.

The advantage of the in-context paradigm is that it provides a simple and direct way to inject knowledge into the model without altering its parameters, making it straightforward to implement. However, this method is constrained by the maximum context length the model can handle and may not enable deep integration of the knowledge into the model's internal reasoning.

2.2 IN-PARAMETER KNOWLEDGE INJECTION PARADIGM

While researchers have extensively explored and clearly defined in-context knowledge augmentation, there has been little investigation into in-parameter knowledge augmentation, and no precise definition for this paradigm has been established. In this paper, we define in-parameter knowledge injection as the process whereby relevant external knowledge is converted into a parametric representation and then incorporated directly into the model's parameters. This approach involves using representation learning techniques to embed the knowledge into the model's parameters, enabling it to utilize this information at a more fundamental level when generating responses.

To achieve this, we define a parameter update function f_ϕ parameterized by ϕ. This function takes the external knowledge $\mathcal{K}$ and the current model parameters θ as inputs to compute a conditional parameter shift $\Delta\theta$. The computation is formalized as:

$$\Delta\theta = f_\phi(\mathcal{K}, \theta), \quad \theta' = \theta + \Delta\theta, \tag{1}$$

where $\Delta\theta$ is the parametric knowledge representation that represents the adjustment to the model parameters necessary to incorporate the external knowledge. The updated parameters θ' integrate this knowledge directly into the model. The updated model $\mathcal{M}'$ with parameters θ' is then used to generate the answer a. This process enables the model to utilize external knowledge at a fundamental level during response generation, enhancing its reasoning capabilities.

The in-parameter paradigm allows the model to deeply integrate knowledge into its parameters, enabling intrinsic use during response generation. This approach overcomes context length limitations and enhances the language model's reasoning capabilities. However, it requires real-time parameter modifications, which, if not properly handled, may introduce additional implementation complexity and computational cost.

3 METHODOLOGY OF IN-PARAMETER KNOWLEDGE INJECTION

Our proposed In-Parameter Knowledge Injection method, illustrated in Figure 2, integrates external knowledge $\mathcal{K}$ directly into the parameters of a pre-trained language model $\mathcal{M}$ by computing Low-Rank Adaptation (LoRA) (Hu et al.) based parametric knowledge representations (the calculation of parametric knowledge representation can be performed offline). This process begins by generating synthetic question-answer (QA) pairs from the provided passages, which are then used to compute

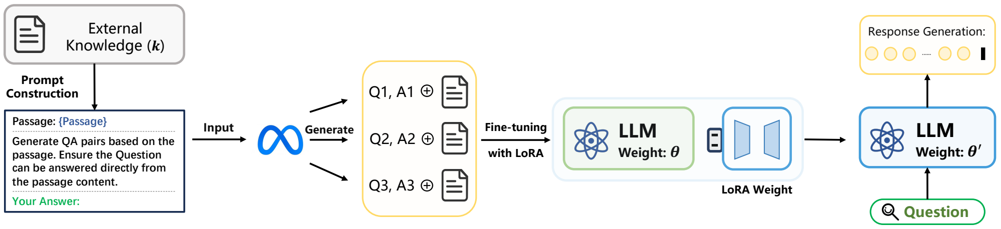


Figure 2: An illustration of our proposed in-parameter knowledge injection framework. The prompt shown is a simplified representation. The detailed prompt is in Appendix A.

LoRA parameters representing the knowledge. These parameters can be directly added to the language model. The process described above consists of two key phases: the Parametric Knowledge Encoding Phase and the Generation Phase. In the following sections, we will provide a detailed explanation of each phase.

3.1 PARAMETRIC KNOWLEDGE ENCODING PHASE

As described in Section 2.2, given a question q and relevant knowledge k, we aim to integrate k into the language model's parametric knowledge and enable the language model to effectively utilize this knowledge for subsequent tasks. The purpose of the Parametric Knowledge Encoding (PKE) Phase is precisely designed to achieve this. In the PKE Phase, we encode the external knowledge $\mathcal{K}$ into parametric knowledge representation that can be directly injected into the model's parameters θ. The process is as follows:

QA Pair Generation Using Auxiliary Language Model To effectively inject the **external knowledge** $\mathcal{K}$ into the language model, we first convert each passage $k \in \mathcal{K}$ into a suitable format for knowledge integration. This process involves generating question-answer (QA) pairs for each passage, which serve as the basis for training the language model to integrate the knowledge.

Specifically, we utilize an auxiliary language model $\mathcal{A}$ to generate QA pairs from each passage. For each $k \in \mathcal{K}$, we apply a structured prompt template $\mathcal{T}$ to format the passage before passing it to $\mathcal{A}$. This prompt template is carefully designed to elicit informative and diverse QA pairs that reflect the core content of the passage (detailed in Appendix A).

Using the prompt template $\mathcal{T}$, the auxiliary language model $\mathcal{A}$ generates a set of QA pairs for each passage k. We aggregate all generated QA pairs into a dataset $\mathcal{D}$:

$$\mathcal{D} = \bigcup_{k \in \mathcal{K}} \left\{ (k, u_i, a_i) \mid i \in 1, 2, \ldots, n \right\}, \tag{2}$$

where $\mathcal{D}$ represents the collection of all tuples (k, u_i, a_i), where each (u_i, a_i) pair corresponds to a specific question and answer derived from passage k, and n is the number of QA pairs we generated, which is a tunable hyperparameter. This dataset serves as the foundation for computing LoRA-based parametric knowledge representations, which consist of additional parameters that can be directly integrated into the language model to internalize the external knowledge represented by $\mathcal{K}$.

LoRA-Based Parametric Knowledge Integration To integrate the external knowledge $\mathcal{K}$ into the language model $\mathcal{M}$, we compute LoRA-based parametric knowledge representations—additional parameters that can be directly added to $\mathcal{M}$ as a form of knowledge embedding.

For each sample (k, u_i, a_i) in $\mathcal{D}$, we construct the input by concatenating the passage k, the question u_i, and the answer a_i into a single sequence. The model is trained using the standard language modeling objective to predict all the tokens in the input sequence based on all preceding tokens. This next-token prediction approach allows the model to learn the contextual relationships between the passage, question, and answer, thereby internalizing the knowledge contained in $\mathcal{K}$.

We employ LoRA to introduce low-rank adaptations to the model's weight matrices. For each weight matrix $W \in \mathbb{R}^{d \times k}$ in the model parameters θ, we introduce low-rank matrices A and B such that:

$$W' = W + \Delta W = W + AB^\top, \tag{3}$$

where $A \in \mathbb{R}^{d \times r}$, $B \in \mathbb{R}^{k \times r}$, and $r \ll \min(d, k)$. These low-rank matrices $\Delta\theta = \{A, B\}$ constitute the parametric knowledge representation that can be directly integrated into the original model $\mathcal{M}$.

The optimization objective is to minimize the negative log-likelihood of the target tokens over the entire input sequence:

$$\min_{\Delta\theta} \mathcal{L}(\theta + \Delta\theta) = - \sum_{(k,u_i,a_i) \in \mathcal{D}} \sum_{t=1}^{T} \log P_{\theta+\Delta\theta}(x_t \mid x_{<t}), \tag{4}$$

where $x = [k; u_i; a_i]$ is the concatenated input sequence, and T is the total number of tokens in the sequence. By computing gradients over the entire input—including the passage, question, and answer—the model effectively learns to internalize the external knowledge.

The resulting LoRA parameters $\Delta\theta$ serve as the parametric knowledge representation, which can be directly added to the original model $\mathcal{M}$ to enhance its performance on tasks requiring the integrated knowledge. **Importantly, this entire process can be performed offline; each passage or a group of passages can be processed in advance to compute their respective parametric representations, eliminating the need for real-time computation during online deployment.**

3.2 GENERATION PHASE

In the Generation Phase, we augment the original model $\mathcal{M}$ with the parametric knowledge representation $\Delta\theta$ to create an updated model $\mathcal{M}'$ with parameters $\theta' = \theta + \Delta\theta$. This updated model is used to generate answers to the question q without providing the knowledge passages $\mathcal{K}$ as input. By internalizing the knowledge into its parameters, the model leverages this information flexibly, enabling deeper reasoning and more informed responses.

3.3 EFFICIENCY COMPARISON BETWEEN IN-CONTEXT AND IN-PARAMETER PARADIGM

The in-context paradigm appends external knowledge directly to the input prompt during inference, increasing the input length and the computational resources required. Specifically, it demands more GPU memory due to the longer sequences processed by the model's self-attention mechanism, whose computational complexity scales quadratically with input length. The in-parameter paradigm involves encoding the parametric knowledge representation. This process consumes time and memory for calculating the representation, **but it can be performed offline**. As a result, the model can simply load these pre-computed parameters for real-time queries without additional processing overhead. During inference, the model operates with input lengths comparable to standard queries without appended knowledge, leading to reduced GPU memory usage and faster inference times.

In summary, while the in-context method saves on training time by eliminating the need for calculating the parametric knowledge representation, it incurs higher inference costs due to longer input sequences and increased GPU memory usage. The in-parameter method front-loads the computational effort during an offline training phase and thus benefits from reduced inference time and memory usage, making it capable of handling real-time queries.

4 EXPERIMENTAL SETUP

4.1 TASKS AND DATASETS

We design a series of progressive experiments tailored to incrementally increase both difficulty and the requisite amount of reasoning needed for the resolution. This structured approach allows us to evaluate the effectiveness of different knowledge injection methods under varying complexities. To be specific, our experiments span from simple fact extraction to complex multi-hop reasoning tasks, using a variety of datasets suited to each task's requirements.

Level 1: Fact Extraction Task Our initial experiments focus on queries that ask about explicit facts directly present in the relevant passage or document without requiring additional reasoning. This is the simplest form of query, where the model's primary task is to locate and extract the relevant information. For this experiment, we use the TriviaQA dataset (Joshi et al., 2017), which consists of question-answer pairs where the answers are explicit facts found within the given passages.

Level 2: Comparative Reasoning Task To introduce a higher level of complexity, we consider tasks that require simple reasoning over information extracted from two documents. Specifically, we focus on **comparison questions**, which involve comparing two or more entities from the same group based on certain attributes. For example, a question might ask, *"Who was born first, Bill Clinton or Donald Trump?"* Answering such questions requires the model to extract relevant facts from multiple documents and perform a comparison. We utilize the **Comparison** subset of the 2WikiMultihopQA (2WQA) dataset (Ho et al., 2020) for this task.

Level 3: Multi-Step Comparative Reasoning Task Further increasing the difficulty, we examine **bridge-comparison questions**, which require an additional reasoning step for each extracted answer before performing the comparison. These questions combine elements of bridge questions with comparison questions. For instance, instead of directly comparing two books, a question might ask, *"Which book has the author born first, Pride and Prejudice or 1984?"* To answer, the model needs to identify the authors of the books and then compare their dates of birth. We use the **Bridge Comparison** subset of the 2WQA dataset for this task.

4.2 EVALUATION METRICS

For all tasks, we evaluate the models based on their ability to provide correct answers. We extract the final answer from the generated output using pattern-matching techniques. The extracted answer is then compared with the reference answer, utilizing methods such as exact match at the answer level, along with token-level measurements of the F1 score. The details of our experimental settings, including the instructions provided to the models and the evaluation protocols, are available in Appendix B.

4.3 IMPLEMENTATION DETAILS

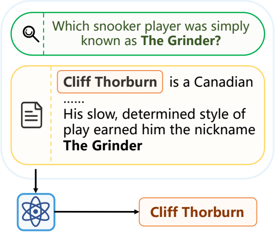
For in-context knowledge injection, we directly concatenate the relevant passages to the prompt template and input the combined text into the language model's context. The specific prompt templates and the designing process are detailed in Appendix D. For in-parameter knowledge injection, we employ GPT-4 as the external model to generate question-answer (QA) pairs based on the passages. The specific configurations, including batch size, epochs, learning rate, and LoRA parameters, are detailed in Appendix B. The experiments utilized the Qwen2.5-1.5B-Instruct (Yang et al., 2024) and LLaMA3.2-1B-Instruct (Meta, 2024) models, with all conducted using PyTorch on 40GB NVIDIA A100 GPUs. The generation settings, including the decoding strategy, and hardware specifics, are detailed in Appendix B.

5 EXPERIMENTAL RESULTS

In this section, we comprehensively evaluate different knowledge injection methods across tasks with varying levels of complexity. Our objective is to understand how each method performs under different reasoning demands and to identify the circumstances under which each method is most effective. The methods compared are:

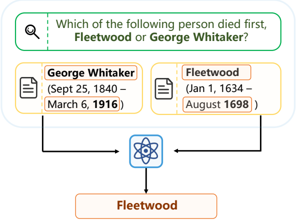
- **In-context (IC)**: Directly inputting the Passage and Question into the language model.
- **IC with QA pairs (IC-QA)**: Inputting the Passage, Question, and the QA pairs generated by the auxiliary language model for knowledge augmentation as introduced in Section 3.1.
- **In-Parameter (IP)**: Injecting knowledge directly into the model's parameters using the method from Section 3.
- **IP & IC**: Combining the In-Parameter method with In-context by both injecting knowledge into the parameters and concatenating the Passage into the input.

Table 1: Comparison of In-context and In-Parameter knowledge injection methods for the Fact Extraction Task, with the best results in bold and second-best results underlined. An accompanying diagram on the left illustrates the task's complexity.


Which snooker player was simply known as The Grinder?
Cliff Thorburn is a Canadian
......
His slow, determined style of play earned him the nickname The Grinder
Cliff Thorburn

	Qwen2.5-1.5B-Instruct		LLaMA3.2-1B-Instruct	
Method	**EM**	**F1**	**EM**	**F1**
In-Context	0.4913	0.6246	0.5896	**0.6979**
IC-QA	**0.5202**	0.6200	**0.6012**	0.6962
In-Parameter	0.2486	0.3685	0.3006	0.3903
IP & IC	0.5087	**0.6604**	0.5260	0.6487

Table 2: Experimental results of In-context and In-Parameter knowledge injection method on the Comparative Reasoning Task. The best results are in bold and the second-best results are underlined. The diagram on the left illustrates the task difficulty.


Which of the following person died first, Fleetwood or George Whitaker?
George Whitaker (Sept 25, 1840 – March 6, 1916)
Fleetwood (Jan 1, 1634 – August 1698)
Fleetwood

	Qwen2.5-1.5B-Instruct		LLaMA3.2-1B-Instruct	
Method	**EM**	**F1**	**EM**	**F1**
In-Context	0.2148	0.3240	0.2517	0.2635
IC-QA	0.1846	0.2426	0.1342	0.1394
In-Parameter	0.3188	0.3826	**0.4765**	**0.5154**
IP & IC	**0.3960**	**0.4473**	0.4732	0.5127

By progressively increasing the difficulty level of the tasks—from simple fact extraction to complex multi-step reasoning—we aim to reveal how each knowledge injection method scales with task complexity and reasoning requirements.

5.1 FACT EXTRACTION TASK

In this initial task, we assess the models' abilities to extract explicit facts directly present in the provided passages without additional reasoning. This task represents the simplest scenario, focusing on straightforward entity extraction. As shown in Table 1, the **In-context** and **IC-QA** methods outperform the **In-Parameter** methods on both the Qwen and LLaMA models. This outcome suggests that when the required information is readily available in the input, providing the passage directly to the model is the most effective approach.

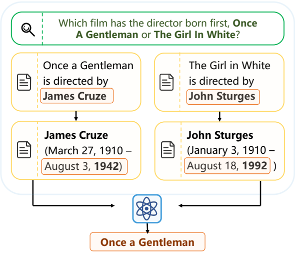
The superior performance of the in-context methods can be attributed to the models' proficiency in understanding and extracting information from the immediate context. On the other hand, the **In-Parameter** methods underperform in this task. One possible explanation is that injecting knowledge into the parameters may introduce unnecessary complexity for simple tasks and might not capture the precise details or could potentially obscure the exact information needed for fact extraction. These findings highlight that for tasks involving direct extraction of information with minimal reasoning, in-context methods are more advantageous.

5.2 COMPARATIVE REASONING TASK

In the next level of complexity, models are required to perform reasoning over information extracted from two documents. This involves comparison questions where the answer is not explicitly stated but must be inferred through basic reasoning. Table 2 reveals a notable shift in performance: the **In-Parameter** and **IP & IC** methods now outperform the in-context approaches on both models. This suggests that as the task requires the integration of multiple pieces of information and reasoning over them, embedding knowledge into the model's parameters becomes more effective.

These findings demonstrate that as complexity and reasoning demands grow, embedding knowledge directly into the model's parameters becomes increasingly beneficial. The In-Parameter method allows the model to internalize external knowledge, enabling deeper integration and more flexi-

Table 3: Performance comparison of In-context and In-Parameter knowledge injection methods on the Multi-Step Comparative Reasoning Task. The best results are in bold and the second-best results are underlined. The diagram on the left illustrates the task difficulty.


Which film has the director born first, **Once A Gentleman** or **The Girl In White**?
Once a Gentleman is directed by **James Cruze**
The Girl in White is directed by **John Sturges**
James Cruze (March 27, 1910 – August 3, **1942**)
John Sturges (January 3, 1910 – August 18, **1992**)
Once a Gentleman

	Qwen2.5-1.5B-Instruct		LLaMA3.2-1B-Instruct	
Method	**EM**	**F1** q	**EM**	**F1**
In-Context	0.2617	0.3578	0.3188	0.3483
IC-QA	0.2181	0.2545	0.2785	0.2969
In-Parameter	**0.3557**	**0.3964**	0.3758	0.4102
IP & IC	0.3289	0.3493	**0.4027**	**0.4278**

ble reasoning across multiple pieces of information. Moreover, combining In-Parameter with In-Context knowledge further enhances performance, indicating that parameter-injected knowledge supplemented by contextual information yields better outcomes.

5.3 MULTI-STEP COMPARATIVE REASONING TASK

At the highest level of complexity, the models tackle multi-step comparative reasoning tasks that require chaining several inference steps before arriving at the answer. This involves not only extracting information but also performing sequential reasoning over that information. As shown in Table 3, the **In-Parameter** and **IP & IC** methods continue to outperform the in-context methods, with the performance gap widening compared to the previous task. This trend underscores the increasing efficacy of in-parameter knowledge injection as the reasoning demands escalate.

A possible reason for this is that the in-parameter methods enable the models to handle the complexity of multi-hop reasoning more effectively. By embedding knowledge into the parameters, the models develop richer internal representations that support complex inferential chains. This internalization allows for more sophisticated reasoning. The combined **IP & IC** method often yields the best performance, suggesting a synergistic effect. Providing the model with both internalized knowledge and contextual information may facilitate reasoning by offering multiple avenues for accessing and processing the necessary data. In contrast, the in-context methods face some challenges in this task. The necessity to conduct multiple reasoning steps within the input context likely overwhelms the models, leading to diminished performance. This highlights the limitations of relying solely on in-context information for complex reasoning tasks.

5.4 ANALYSIS

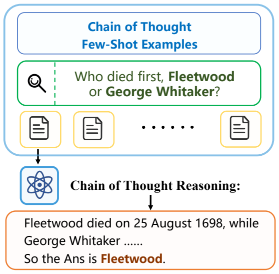
The experiments indicate a clear trend: in-parameter knowledge injection becomes more effective as task complexity and reasoning demands increase, enhancing the model's deeper reasoning capabilities. In-context methods perform better in simple fact extraction tasks as models can readily utilize the provided context. However, for higher-order reasoning, in-context approaches are limited by the capacity to reason based on the injected knowledge. In such cases, in-parameter methods provide an advantage by embedding knowledge internally, enabling deeper reasoning based on its parametric knowledge. These findings suggest that selecting a knowledge injection method based on the task complexity can significantly enhance model performance.

6 ABLATION STUDY

6.1 INFLUENCE OF FEW-SHOT CHAIN-OF-THOUGHT EXAMPLES

To evaluate whether incorporating few-shot chain-of-thought (CoT) (Wei et al., 2023) examples affects our conclusions, we introduced CoT into the prompts for the Comparative Reasoning Task and the Multi-Step Comparative Reasoning Task. The experimental results are shown in Table 4.

Table 4: Comparison of performance with and without few-shot chain-of-thought (CoT) prompting. The best and second-best results are highlighted in bold and underlined, respectively. The left figure illustrates the setting. The models used are Qwen2.5-1.5B-Instruct and LLaMA3.2-1B-Instruct.


Chain of Thought
Few-Shot Examples
Who died first, Fleetwood or George Whitaker?
Chain of Thought Reasoning:
Fleetwood died on 25 August 1698, while George Whitaker ......
So the Ans is Fleetwood.

	Comparative Reasoning		Multi-Setp Comparison	
Method	**Qwen**	**LLaMA**	**Qwen**	**LLaMA**
In-Context	0.3154	0.3960	0.4094	0.2013
IC-QA	0.2987	0.3020	**0.4195**	0.2483
In-Parameter	0.3456	0.4497	0.4027	**0.4094**
IP & IC	**0.3859**	**0.4765**	0.3993	0.3523

The In-Parameter and IP & IC methods continued to outperform the in-context approaches. This consistency suggests that embedding knowledge directly into the model's parameters is inherently effective for complex reasoning tasks, regardless of the presence of CoT examples in the prompts. These findings indicate that our earlier conclusions about the superiority of in-parameter knowledge injection are robust and not significantly influenced by the addition of CoT examples. The models' ability to handle intricate reasoning appears to rely more on internalized knowledge than on explicit reasoning cues provided during inference.

6.2 PERFORMANCE ON LARGER MODEL

To explore whether the observed trends persist with larger models, we conducted experiments on the comparison dataset using the LLaMA3-8B-Instruct model, which has a significantly larger parameter count compared to the models used in previous experiments.

As shown in Figure 3 these results are consistent with our earlier findings. The **In-Parameter** and **IP & IC** methods outperform the in-context approaches, reinforcing the effectiveness of embedding knowledge into the model's parameters for tasks requiring reasoning over multiple documents. Notably, the **IP & IC** method achieves the highest performance, suggesting that combining internalized knowledge with contextual information continues to provide benefits even in larger models.

6.3 EFFECT OF QA PAIR GENERATION

To evaluate the effectiveness of the QA pair generation step in our In-Parameter Knowledge Injection method, we conducted an ablation study comparing two training strategies:

- Passage-Only Training: The language model is fine-tuned directly on the knowledge passages $\mathcal{K}$ using the standard language modeling objective, without generating QA pairs.
- Passage with Generated QA Training: Our proposed method, where an auxiliary language model generates QA pairs from each passage, and the model is fine-tuned on the concatenated sequences of passage, question, and answer.

As shown in Table 5, incorporating generated QA pairs into the training process significantly enhances the model's performance. The model trained with QA pairs outperforms the passage-only model by a substantial margin across all evaluation metrics, demonstrating the effectiveness of our approach. The notable performance improvement can be attributed to the explicit question-answering context provided by the generated QA pairs. Training on these pairs enables the model to better internalize and organize the external knowledge, learning not just the memorize the content of the passages but also how to apply this knowledge to respond to specific queries. In contrast, passage-only training lacks this targeted learning mechanism, leading to less effective knowledge integration and application.

Figure 3: EM score of LLaMA3-8B-Instruct on the comparative reasoning task.

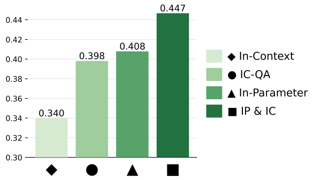


Table 5: Ablation study results comparing Passage-Only Training and Passage with Generated QA Training. The best results are in bold.

		Qwen2.5-1.5B-Instruct		LLaMA3.2-1B-Instruct	
Method	**QA**	**EM**	**F1**	**EM**	**F1**
In-Parameter	**w QA**	**0.3188**	**0.3826**	**0.4765**	**0.5154**
	w/o QA	0.0168	0.0773	0.3926	0.4246
IP & IC	**w QA**	**0.3960**	**0.4473**	**0.4732**	**0.5127**
	w/o QA	0.1107	0.1926	0.2819	0.3123

7 RELATED WORKS

In-Context Knowledge Injection In-context knowledge injection is a prevalent method for augmenting language models with external information by appending relevant passages directly to the input context. This approach has been widely used in tasks such as reading comprehension and Retrieval-Augmented Generation (RAG), where models generate responses based on both the prompt and the injected knowledge (Lewis et al., 2020; Zhou et al., 2024; Su et al., 2024). RAG systems typically retrieve relevant documents and incorporate them into the input to enhance performance on specific tasks. To improve the efficacy of in-context knowledge injection, some studies have focused on optimizing prompts and instructions. For example, Trivedi et al. (2022) introduced IR-CoT, which investigates how to design prompts and few-shot examples to effectively integrate knowledge into the context, thereby enhancing the model's reasoning capabilities over the injected information. To address the limited context window of language models, various context compression techniques have been proposed to mitigate constraints on the amount of external knowledge that can be included (Ge et al., 2023; Verma, 2024).

In-Parameter Knowledge Injection In contrast, in-parameter knowledge injection embeds external knowledge directly into the model's parameters, offering the potential to incorporate more extensive and nuanced information without the constraints of input length. This approach is relatively underexplored, with the most closely related areas being knowledge editing and continued pre-training. Knowledge editing methods, such as Knowledge Neurons (Dai et al., 2021), Rank-One Model Editing (Meng et al., 2022), and Self-Edit (Liu et al., 2024), permanently modify language models to incorporate new information, typically addressing specific facts or entities. These approaches are not designed for the temporary integration of extensive external knowledge tailored to specific tasks. Continued pre-training increases a model's knowledge through further training on extra data, yet it requires substantial resources and time. To alleviate this, parameter-efficient fine-tuning (PEFT) techniques like LoRA (Hu et al.), Adapter (Houlsby et al., 2019), and Prefix-Tuning (Li & Liang, 2021) serve as alternatives. These methods allow efficient integration of knowledge with minimal updates to model parameters. For instance, LoRA adjusts model weights with low-rank updates, and Prefix-Tuning enhances input sequences with a learnable prefix. **Although these approaches improve task performance efficiently, they mainly focus on task adaptation rather than quick, query-specific knowledge updates.**

In contrast, our method temporarily injects external knowledge directly into the model's parameters by encoding the knowledge into parametric knowledge representation. Much like directly appending a passage to the input context, our method does not permanently alter the parameters of the LLM. Instead, the passage related to a specific query is temporarily embedded into the parameters, serving only that query.

8 CONCLUSION

In this paper, we proposed an effective in-parameter knowledge augmentation method, directly integrating external knowledge into generative language model parameters via representation learning. Our experiments compared in-context and in-parameter methods across diverse tasks. Results show that in-parameter knowledge injection excels in reasoning-intensive scenarios, such as advanced comprehension and multi-document inference, due to deeper knowledge integration. In contrast, the in-context method is better suited for simple tasks where answers are directly extractable. The selection between these two methods in practical applications depends on a trade-off: in-parameter offers performance gains for complex reasoning, while in-context minimizes latency for simpler tasks. Practitioners should decide based on specific application needs.

9 REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our results, we have thoroughly documented all settings and details within the paper. Moreover, we have meticulously organized and open-sourced all the code, data, and models used in this study. These resources are available at the following anonymous GitHub link: `https://anonymous.4open.science/r/In-parameter-Knowledge-Injection/`.

REFERENCES

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. *arXiv preprint arXiv:2104.08696*, 2021.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

Tao Ge, Jing Hu, Lei Wang, Xun Wang, Si-Qing Chen, and Furu Wei. In-context autoencoder for context compression in a large language model. *arXiv preprint arXiv:2307.06945*, 2023.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. *arXiv preprint arXiv:2011.01060*, 2020.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pp. 2790–2799. PMLR, 2019.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. Same task, more tokens: the impact of input length on the reasoning performance of large language models. *arXiv preprint arXiv:2402.14848*, 2024.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474, 2020.

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. Long-context llms struggle with long in-context learning. *arXiv preprint arXiv:2404.02060*, 2024a.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

Xiaopeng Li, Shasha Li, Shezheng Song, Jing Yang, Jun Ma, and Jie Yu. Pmet: Precise model editing in a transformer, 2024b. URL `https://arxiv.org/abs/2308.08742`.

Jiateng Liu, Pengfei Yu, Yuji Zhang, Sha Li, Zixuan Zhang, and Heng Ji. Evedit: Event-based knowledge editing with deductive editing boundaries, 2024.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022.

Meta. Llama 3.2 1b instruct, 2024. URL `https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct`. Accessed: 2024-09.

Neel Nanda, Senthooran Rajamanoharan, János Kramár, and Rohin Shah. Fact finding: Attempting to reverse-engineer factual recall on the neuron level, December 2023. URL `https://www.lesswrong.com/posts/iGuwZTHWb6DFY3sKB/fact-finding-attempting-to-reverse-engineer-factual-recall`.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. DRAGIN: Dynamic retrieval augmented generation based on the real-time information needs of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12991–13013, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.702. URL `https://aclanthology.org/2024.acl-long.702`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *arXiv preprint arXiv:2212.10509*, 2022.

Sourav Verma. Contextual compression in retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2409.13385*, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL `https://arxiv.org/abs/2201.11903`.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Yujia Zhou, Zheng Liu, Jiajie Jin, Jian-Yun Nie, and Zhicheng Dou. Metacognitive retrieval-augmented large language models. In *WWW*, pp. 1453–1463. ACM, 2024.

A PROMPT TEMPLATE FOR QA GENERATION

In order to systematically generate high-quality question-answer (QA) pairs from each knowledge passage, we designed a specialized prompt template for the auxiliary language model. The primary objective of this prompt is to ensure that the generated QA pairs are both informative and diverse, accurately reflecting the essential content of the original passage. This structured approach facilitates effective knowledge integration by providing the language model with clear and consistent training data.

A.1 DESIGN CONSIDERATIONS

Several key factors influenced the design of the prompt template:

- **Clarity and Specificity**: The prompt explicitly instructs the model to generate three distinct questions, each answerable using the passage.
- **Structured Output Format**: By specifying the exact format for the questions and answers, including fields such as “question”, “answer”, and “full_answer”, we facilitate automated parsing and processing of the generated data. We used a simple example to help the model understand the format we need.
- **Quality Control**: Requiring that each question be answerable using the provided passage helps maintain the relevance and accuracy of the QA pairs, preventing the introduction of extraneous or speculative information. This clarity helps in minimizing ambiguity and ensures that the generated QA pairs are directly relevant to the source material.

A.2 PROMPT TEMPLATE

The following prompt template is used to guide the auxiliary language model in generating QA pairs:

Prompt Template for QA Generation

```
I will provide a passage of text, and you need to generate three different questions
based on the content of this passage. Each question should be answerable using the
information provided in the passage. Additionally, please provide an appropriate
answer for each question derived from the passage.
You need to generate the question and answer in the following format:
[
    {
        “question”: “What is the capital of France?”,
        “answer”: “Paris”,
        “full_answer”: “The capital of France is Paris.”
    },
]
This list should have at least 3 elements. You only need to output this list in the
above format.
Passage:
{passage}
```

When applying the prompt template, the passage is dynamically inserted into the designated placeholder, ensuring that each QA generation task is contextually tied to its corresponding knowledge excerpt. The auxiliary language model processes this prompt to produce a structured list of simple single-hop QA pairs, adhering strictly to the specified format.

B EXPERIMENT DETAILS

This appendix provides a comprehensive overview of the models used, the experimental setting, and the evaluation methodologies employed in our study.

B.1 MODEL DESCRIPTIONS

In our experiments, we utilized two language models Qwen2.5-1.5B-Instruct and LLaMA3.2-1B-Instruct. Thes instruction-tuned models excel in dialogue applications, offering enhanced performance for complex tasks.

- **Qwen2.5-1.5B-Instruct** (Yang et al., 2024): Qwen2.5-1.5B-Instruct is part of the latest Qwen2.5 series of multilingual large language models, featuring 1.5 billion parameters. This instruction-tuned, text-only model is optimized for multilingual dialogue use cases, including agentic retrieval and summarization tasks. Qwen2.5-1.5B-Instruct supports long-context handling of up to 128K tokens and is fine-tuned using supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) to enhance helpfulness and safety.
- **LLaMA3.2-1B-Instruct** (Meta, 2024): LLaMA3.2-1B-Instruct model is an instruction-tuned, text-only variant optimized for multilingual dialogue use cases, including agentic retrieval and summarization tasks. Architecturally, LLaMA3.2 employs an optimized transformer-based auto-regressive framework, ensuring efficient and scalable language generation capabilities. The instruction-tuned versions leverage supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) to better align with human preferences for helpfulness and safety.

B.2 EXPERIMENTAL SETTING

Our experimental framework comprises two primary datasets: TriviaQA and 2WikiMultihopQA. Below, we provide detailed descriptions of the setting and procedures applied to each.

B.2.1 2WIKIMULTIHOPQA

We employed two data types from the 2WikiMultihopQA dataset: Comparison and Bridge Comparison. For each data type, the first 298 questions were selected to generate responses for evaluation during the main experiment. The ablation study experiment for LLaMA3-8B-Instruct utilizes the first 103 questions.

For every question within these categories, we extracted the pertinent information and created three question-answer pairs for each extracted segment. To enhance the model's ability to recall and effectively utilize passage information, we organized the training data for in-parameter training as follows:

- **(Q, P, A) Prompts**: Two QA pairs combine the question (Q), passage (P), and answer (A), encouraging the model to associate the passage with the corresponding answer.
- **(Q, A) Prompts**: One QA pair includes only the question and answer, facilitating the model's ability to generate answers based on the learned passage content.

B.2.2 TRIVIAQA

We employed the Wikipedia development set from TriviaQA, where all relevant documents are sourced from Wikipedia. To reduce the complexity of reading comprehension for the model, we implemented a filtering process that retained only passages containing 3,000 tokens or fewer for each question. Initially, the dataset included 308 questions; after filtering, 173 questions remained, each containing at least one relevant piece of information within the length limit. These filtered questions were used for our experiments.

For every question within these categories, we extracted the pertinent information and created six question-answer pairs for each extracted segment. Following the aforementioned format in 2WikiMultihopQA, each question includes three (Q, P, A) prompts and three (Q, A) prompts to enhance the model's ability to effectively recall and utilize passage information.

B.3 EVALUATION DETAILS

To assess the performance of our models, we employed two primary evaluation metrics: Exact Match (EM) and F1 Score.

For tasks that require direct answer generation, we incorporate the prompt phrase "The answer is" within the input. This encourages the model to produce the answer immediately following this phrase, ensuring consistency with the expected output format.

For tasks that involve reasoning and necessitate the generation of a Chain-of-Thought (CoT), we provide eight few-shot examples that utilize CoT. Each few-shot example concludes with the phrase "So, the answer is `xxx`", which guides the model to follow this pattern and facilitates accurate answer matching.

B.4 TRAINING PARAMETER CONFIGURATION

The training parameters and LoRA parameters used in our In-Parameter Knowledge Injection method are on Table 6.

Table 6: Configuration of Training Parameters

Parameter	**Value**
Number of Training Epochs	3
Learning Rate	3×10^{-4}
LoRA Alpha	32
LoRA Dropout	0.01
LoRA Rank	Qwen2.5-1.5B-Instruct: 2 LLaMA3.2-1B-Instruct: 2 LLaMA3-8B-Instruct: 16

C MORE ABLATION STUDY RESULTS

C.1 EFFECT OF LORA PARAMETERS

To investigate the impact of LoRA parameters, we conducted an ablation study using the LLaMA3.2-1B-Instruct model on the first 103 questions of the Comparison dataset. We systematically varied the rank, α and dropout rates, as detailed in Table 7, to assess their effects on performance.

The experimental results indicate that changes in LoRA parameters have little influence on the model's performance. For smaller models, a lower rank typically leads to better performance. The alpha parameter has a relatively minor impact, while a lower dropout rate also contributes to improved performance.

Table 7: Ablation study results comparing different rank, α, and dropout configurations. The best results are in bold.

Rank	α	**Dropout**	**EM**	**F1**	**Prec.**	**Recall**
2	16	0.01	0.4854	0.4999	0.4977	0.5100
2	32	0.01	**0.5631**	**0.5994**	**0.5959**	**0.6081**
2	32	0.05	0.4175	0.4244	0.4236	0.4291
4	32	0.01	0.5243	0.5267	0.5275	0.5262
8	32	0.01	0.4563	0.4636	0.4628	0.4680
16	32	0.01	0.4466	0.4555	0.4547	0.4583

C.2 EFFECT OF THE NUMBER OF TRAINING EPOCHS AND KNOWLEDGE-AUGMENTED QAS

To investigate the influence of training parameters on model performance, we conducted an ablation study using the LLaMA3.2-1B-Instruct model on the first 103 questions of the Comparison dataset. Specifically, we examined the effects of varying the number of training epochs (#Epoch) and the number of knowledge-augmented question-answer pairs (#QA). The performance is presented in Tables 8 and 9.

Table 8: Ablation study results comparing different numbers of training epochs. The best results are in bold.

#Epoch	EM	F1	Prec.	Recall
1	0.5534	0.5733	0.5744	0.5748
3	**0.5631**	**0.5994**	**0.5959**	**0.6081**
5	0.3981	0.4099	0.4091	0.4146
7	0.3592	0.3893	0.3859	0.4008

The results in Table 8 indicate that three training epochs yield the best performance across all metrics. Training for a single epoch also achieves relatively strong results, but increasing the number of epochs beyond three leads to a significant decline in performance. Similarly, Table 9 shows that increasing the number of knowledge-augmented QAs up to three substantially improves the model's performance, with the optimal results observed at three QAs. However, adding more than three QAs results in decreased performance.

These results suggest that excessive training may lead to overfitting or weaken the model's ability to generalize effectively, ultimately causing it to lose its ability to answer questions accurately.

Table 9: Ablation study results comparing different numbers of knowledge-augmented QA. The best results are in bold.

#QA	**EM**	**F1**	**Prec.**	**Recall**
1	0.4272	0.4387	0.4384	0.4453
2	0.4757	0.4911	0.4896	0.4947
3	**0.5631**	**0.5994**	**0.5959**	**0.6081**
4	0.4369	0.4523	0.4502	0.4615
5	0.4272	0.4535	0.4503	0.4647
6	0.4175	0.4199	0.4207	0.4194

D PROMPT DETAILS

For methods that perform In-Context Learning solely on passages, we utilize the following prompt:

Prompt Template for In-Context Learning with Passages Only

user:

You should answer the question by referring to the knowledge provided below and integrating your own knowledge.
Passage 1: Blind Shaft is a 2003 film about a pair of brutal con artists operating in the illegal coal mines of present- day northern China. The film was written and directed by Li Yang, and is based on Chinese writer Liu Qingbang’s short novel” Shen MuSacred Wood”).
Passage 2: The Mask of Fu Manchu is a 1932 pre-Code adventure film directed by Charles Brabin. It was written by Irene Kuhn, Edgar Allan Woolf and John Willard based on the 1932 novel of the same name by Sax Rohmer. Starring Boris Karloff as Fu Manchu, and featuring Myrna Loy as his depraved daughter, the movie revolves around Fu Manchu’s quest for the golden sword and mask of Genghis Khan. Lewis Stone plays his nemesis. Dr. Petrie is absent from this film.

Question: Which film came out first, Blind Shaft or The Mask Of Fu Manchu?

assistant:

The answer is

For the (Q, P, A) part of the In-Parameter training data, the data is the prompt above with the answer added at the end.

For the part that does not use knowledge, we use the following prompt:

Prompt Template for In-Context Learning with Passages Only

user:

You should answer the question by referring to the knowledge provided below and integrating your own knowledge.

Question: Which film came out first, Blind Shaft or The Mask Of Fu Manchu?

assistant:

The answer is

For the (Q, A) part of the In-Parameter training data, the data is the prompt above with the answer added at the end.

For methods that perform In-Context Learning on both passages and QA-generated knowledge, we use the following prompt:

Prompt Template for In-Context Learning with Passages and Knowledge-Augmented QA

user:

You should answer the question by referring to the knowledge provided below and integrating your own knowledge.
Passage 1: Blind Shaft is a 2003 film about a pair of brutal con artists operating in the illegal coal mines of present-day northern China. The film was written and directed by Li Yang, and is based on Chinese writer Liu Qingbang’s short novel ”Shen Mu (Sacred Wood)”.
Passage 2: The Mask of Fu Manchu is a 1932 pre-Code adventure film directed by Charles Brabin. It was written by Irene Kuhn, Edgar Allan Woolf, and John Willard based on the 1932 novel of the same name by Sax Rohmer. Starring Boris Karloff as Fu Manchu, and featuring Myrna Loy as his depraved daughter, the movie revolves around Fu Manchu’s quest for the golden sword and mask of Genghis Khan. Lewis Stone plays his nemesis. Dr. Petrie is absent from this film.

Here are some questions and answers about the knowledge.
Question: What is the film ”Blind Shaft” about?
Answer: A pair of brutal con artists operating in the illegal coal mines of present-day northern China.
Question: Who wrote and directed the film ”Blind Shaft”?
Answer: Li Yang.
Question: What is the source material for the film ”Blind Shaft”?
Answer: Chinese writer Liu Qingbang’s short novel ”Shen Mu (Sacred Wood)”.
Question: Who directed the film The Mask of Fu Manchu?
Answer: Charles Brabin.
Question: Who played the character Fu Manchu in the film?
Answer: Boris Karloff.
Question: What is Fu Manchu seeking in the movie?
Answer: The golden sword and mask of Genghis Khan.

You need to answer the question: Which film came out first, Blind Shaft or The Mask of Fu Manchu?

assistant:

The answer is

For tasks that require guiding the model to generate Chain-of-Thought (CoT) reasoning, we use the following instruction to provide few-shot examples before the external knowledge:

Prompt Template for CoT fewshot

You should reference the knowledge provided below and combine it with your own knowledge to answer the question. Please follow the format of the example I provided above.
Here are some examples about how to answer the questions.
Question: When did the director of film Hypocrite (Film) die?
Answer: The film Hypocrite was directed by Miguel Morayta. Miguel Morayta died on 19 June 2013. So the answer is 19 June 2013.

Question: Are both Kurram Garhi and Trojkrsti located in the same country?
Answer: Kurram Garhi is located in the country of Pakistan. Trojkrsti is located in the country of Republic of Macedonia. Thus, they are not in the same country. So the answer is no.

Question: Do director of film Coolie No. 1 (1995 Film) and director of film The Sensational Trial have the same nationality?
Answer: Coolie No. 1 (1995 film) was directed by David Dhawan. The Sensational Trial was directed by Karl Freund. David Dhawan's nationality is India. Karl Freund's nationality is Germany. Thus, they do not have the same nationality. So the answer is no.

Question: Who is Boraqchin (Wife Of Ögedei)'s father-in-law?
Answer: Boraqchin is married to Ögedei Khan. Ögedei Khan's father is Genghis Khan. Thus, Boraqchin's father-in-law is Genghis Khan. So the answer is Genghis Khan.

Question: Who was born first out of Martin Hodge and Ivania Martinich?
Answer: Martin Hodge was born on 4 February 1959. Ivania Martinich was born on 25 July 1995. Thus, Martin Hodge was born first. So the answer is Martin Hodge.

Question: When did the director of film Laughter In Hell die?
Answer: The film Laughter In Hell was directed by Edward L. Cahn. Edward L. Cahn died on August 25, 1963. So the answer is August 25, 1963.

Question: Which film has the director died later, The Gal Who Took the West or Twenty Plus Two?
Answer: The film Twenty Plus Two was directed by Joseph M. Newman. The Gal Who Took the West was directed by Frederick de Cordova. Joseph M. Newman died on January 23, 2006. Fred de Cordova died on September 15, 2001. Thus, the person to die later from the two is Twenty Plus Two. So the answer is Twenty Plus Two.

Question: Who is the grandchild of Krishna Shah (Nepalese Royal)?
Answer: Krishna Shah has a child named Rudra Shah. Rudra Shah has a child named Prithvipati Shah. Thus, Krishna Shah has a grandchild named Prithvipati Shah. So the answer is Prithvipati Shah.

The passages and knowledge-augmented question-answer pairs follow the same format as above. In the generation part, we guide the model to think step-by-step.

Prompt Template for CoT generation

user:
......
Let's think step by step. Answer the questions in the same format as above.
Question: Which film came out first, Blind Shaft or The Mask Of Fu Manchu?

assistant:

Answer: