Limited Out-of-Context Knowledge Reasoning in Large Language Models

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have demonstrated strong capabilities as knowledge bases and significant in-context reasoning capabilities. However, previous work challenges their out-of-context reasoning ability, i.e., the ability to infer information from their training data, instead of from the context or prompt. This paper focuses on a significant facet of out-of-context reasoning: Out-of-Context Knowledge Reasoning (OCKR), which is to combine multiple knowledge to infer new knowledge. We designed a synthetic dataset with seven representative OCKR tasks to systematically assess the OCKR capabilities of LLMs. Using this dataset, we evaluated the LLaMA2-13B-chat model and discovered that its proficiency in this aspect is limited, regardless of whether the knowledge is trained in a separate or adjacent training settings. Moreover, training the model to reason with complete reasoning data did not result in significant improvement. Training the model to perform explicit knowledge retrieval helps in only one of the tasks, indicating that the model's limited OCKR capabilities are due to difficulties in retrieving relevant knowledge. Furthermore, we treat cross-lingual knowledge transfer as a distinct form of OCKR, and evaluate this ability. Our results show that the evaluated model also exhibits limited ability in transferring knowledge across languages.

1 Introduction

In the realm of in-context learning, LLMs not only demonstrate significant reasoning capabilities (Kojima et al., 2022; Yao et al., 2023; Besta et al., 2023) but also concurrently exhibit expertise as knowledge bases in various academic and professional domains, including science, history, law, and finance (Petroni et al., 2019; Wei et al., 2023; AlKhamissi et al., 2022). However, it is unclear whether their reasoning ability is limited to in-context scenarios, or they can also perform out-of-context reasoning, which, as defined by previous studies (Berglund et al., 2023a), is "to recall facts learned in training and use them at test time, despite these facts not being directly related to the test-time prompt." Berglund et al. (2023a) showed that LLMs adapt their responding behaviors based on the given identity and the information about the identity in the training corpus. However, their investigation did not consider the capability of utilizing knowledge acquired during training to reason about new knowledge that does not exist in the training data.

For instance, if an LLM knows from the training data that *Joe Biden was born in 1942* and *Stephen William Hawking shares the same birth year with Joe Biden*, can it infer *Hawking's birth year* as 1942 without having been directly trained on this specific fact? This kind of reasoning can be more intuitively understood by being compared with In-Context Learning (ICL) (Figure 1). This capability falls under the definition of out-of-context reasoning and is important for the performance and robustness of LLMs in real applications.

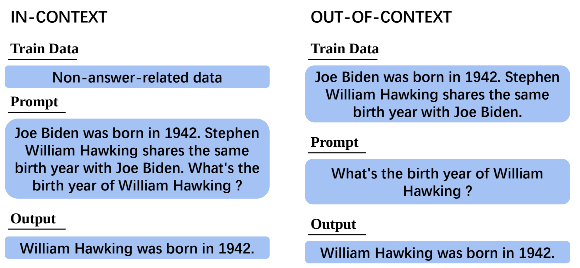

Figure 1: In-Context vs Out-of-Context. In the In-Context scenario, the relevant data is provided in the prompt to allow the model to infer the answer. In the Out-of-Context scenario, the relevant data is included directly in the training data, and the model is then asked to infer the answer based on this training.

This paper proposes the investigation of Out-of-Context Knowledge Reasoning (OCKR), a vital component of out-of-context reasoning. We

Combination	Examples of T_1 and T_2	Feasibility and possible $\bar{T}$
$A \wedge A \to A$	(x, birth_year, 2000) (y, birth_year, 2000)	No, cannot infer new meaningful attributes
$A \wedge A \to R$	(x, birth_year, 2000) (y, birth_year, 2000)	Yes, e.g.: (x, birth_year_equals, y)
$A \wedge R \to A$ ($R \wedge A \to A$)	(x, birth_year, 2000) (x, birth_year_equals, y)	Yes, e.g.: (y, birth_year, 2000)
$A \wedge R \to R$ ($R \wedge A \to R$)	(x, birth_year, 2000) (x, birth_year_equals, y)	No, cannot infer new meaningful relationships
$R \wedge R \to A$	(x, birth_year_equals, y) (x, birth_year_equals, z)	No, pure relationships cannot infer attributes
$R \wedge R \to R$	(x, birth_year_equals, y) (x, birth_year_equals, z)	Yes, e.g.: (y, birth_year_equals, z)

Table 1: Feasibility analysis of all possible combinations for the reasoning patterns. x, y, and z denote specific entities involved in the training process. Considering the interchangeability of T_1 and T_2, redundant combinations are eliminated. For $A \wedge A \to A$ and $A \wedge R \to R$, it is difficult to derive meaningful new knowledge without borrowing other external knowledge. For $R \wedge R \to A$, attributes cannot be inferred from pure relationships. Consequently, we identify $A \wedge A \to R$, $A \wedge R \to A$, and $R \wedge R \to R$ as viable knowledge reasoning patterns.

propose a formal definition of the problem to facilitate discussion. We discuss and design 7 related tasks covering reasoning over different kind of knowledge, such as attributes (A) and relations (R), and construct corresponding datasets to systematically evaluate the OCKR abilities. The evaluation on several open-source LLMs, e.g. LLaMA2-13B-CHAT (Touvron et al., 2023), Baichuan2-13B-CHAT (Yang et al., 2023), Pythia-12B (Biderman et al., 2023), shows that these LLMs have very limited OCKR ability.

Intuitively, new knowledge can emerge during the training or inference phase. We also conduct experiments to assist the LLMs to perform OCKR in different phases, which serve as in-depth analyses for the potential difficulties of reasoning. In the training phase, we merge related knowledge into adjacent text, which may be easier for reasoning. In the inference phase, we train the LLMs to learn the reasoning pattern, or provide them with chain-of-thought (COT) prompt, explicitly retrieving and applying the knowledge. We also study the cross-lingual OCKR as a special case.

Taking LLaMA2-13B-CHAT (Touvron et al., 2023) as a representative model for the above analysis, our main findings are:

- The model shows limited OCKR ability even with knowledge occurs adjacently during training.
- Training the model with reasoning examples does not lead to significant improvement, suggesting that enhancing reasoning ability in general is insufficient for effective OCKR.
- With the help of CoT, the model achieves over 90% accuracy in one task ($A \wedge A \to R$) but does not surpass the random level in other two tasks ($A \wedge R \to A$ and $R \wedge R \to R$). This indicates that the model can effectively retrieve attribute knowledge but struggles with correctly retrieving relational knowledge, which might be a limiting factor in OCKR.
- In both the Separate and Adjacent settings, the performance in cross-lingual scenarios surpasses that of the monolingual $A \wedge R \to A$. However, the overall performance are still weak.

2 Problem Definition

2.1 OCKR Problems

An example of OCKR can be formally represented as:

$$T_1 \wedge T_2 \wedge \ldots \wedge T_n \to \bar{T} \quad (n \geq 1) \qquad (1)$$

where $T_1, T_2, \ldots, T_n$ denotes knowledge in training data; $\bar{T}$ denotes knowledge not in the training data; with the constraint that $T_1, T_2, \ldots, T_n$ are sufficient to imply $\bar{T}$. If a given model trained on $T_1, T_2, \ldots, T_n$ can correctly answer question about $\bar{T}$, , we say that the model has n-ary OCKR ability, i.e. the model can infer $\bar{T}$ from $T_1, T_2, \ldots, T_n$.

In this paper, we focus on the binary OCKR case where n = 2, i.e., $T_1 \wedge T_2 \rightarrow \bar{T}$, which is the simplest case that allows knowledge to be reasoned between different entities.

The knowledge considered in this study falls into two categories according to the knowledge graph taxonomy: Attributes (A) and Relations (R). They are involved with entities in triplets, i.e., (Entity, Attribute, Value) for attributes, (Entity, Relation, Entity) for relations (Kejriwal et al., 2021).

By using A and R as the known knowledge and infer new knowledge, six potential combinations can be enumerated. Among them, only three combinations can be aligned with feasible knowledge reasoning patterns. They are: Attribute $\wedge$ Attribute $\rightarrow$ Relationship ($A \wedge A \rightarrow R$), Attribute $\wedge$ Relationship $\rightarrow$ Attribute ($A \wedge R \rightarrow A$), and Relationship $\wedge$ Relationship $\rightarrow$ Relationship ($R \wedge R \rightarrow R$). See Table 1 for details. Thus, we choose these three types of reason tasks for further study.

2.2 Dataset Design

This paper introduces the Inference Dataset for OCKR (ID-OCKR). The dataset encompasses seven subsets, including the three knowledge reasoning patterns, each presented at both simple and hard levels, along with a subset specifically designed for evaluating cross-lingual capabilities. See Table 2 for details.

Knowledge Assessing the model's OCKR capabilities is non-trivial, because it is not easy to discriminate whether the knowledge is derived from the training data or actually exists in the training data. Furthermore, the LLMs language ability has a huge impact on its performance in different benchmarks. Therefore, it is essential to create a fictional set of knowledge that doesn't rely on knowledge of existing facts, and minimize the language barrier in understanding the knowledge.

Therefore, we choose a very simple attribute, i.e. the year of birth, and some simple relations based on this single attribute, i.e. birth in the same year, birth year greater (i.e. older), one year older, etc, to avoid complex knowledge understanding. For adding a little challenge in the reasoning process, there are two levels of tasks. For the simple level of the task, the relation is only about the equivalence of the attributes; while for the hard level, the relation may need a numerical comparison or calculation of attributes.

Cross-lingual Task The motivation for constructing a cross-lingual dataset stems from our recognition of translation as a unique relation type that links an entity to its translated counterpart. This allows us to conceptualize cross-lingual knowledge transfer as involving three components: attribute knowledge in English (A), translation knowledge (relation between English entity and the corresponding entity in another language, i.e. R), and attribute knowledge the other language (denoted as A). Thus, the cross-lingual scenario can be formally represented as a special form of $A \wedge R \rightarrow A$.

To capture a wide range of linguistic diversity, we selected nine languages based on their widespread use and diverse linguistic families: German (de), French (fr), Italian (it), Russian (ru), Polish (pl), Arabic (ar), Hebrew (he), Chinese (zh), and Japanese (ja). See Table 3 for more details.

2.3 Datasets Construction

We utilize GPT-4(Achiam et al., 2023) to create fictitious entities for our dataset. The name of the entities are constructed using fantastical words, such as “ReverentDawn", to ensure they are rarely found in the original corpus.

For each knowledge template listed in Table 2, 420 entities are created with fictitious name and randomly assigned attributes (i.e. random birth years between 1991 and 2010). The relation between entities are decided based on the attributes of the entities. The datasets scale is in the thousands.

We also utilize GPT-4 to generate 10 text templates for each attribute and relations. For constructing the training set, and all 10 templates are used for describing the knowledge; while for the test set, only one of the text templates are used to determine the result. Examples of the generated instances are presented in Table 9 in Appendix C. For additional details on the organization of datasets, please refer to Appendix D.

3 Methodology

3.1 Evaluation of OCKR

We perform training and test using the ID-OCKR dataset. For training, we fine-tune the LLMs so that the accuracy of responses to knowledge triples in the training set exceeds 90% to ensure the LLM learns the knowledge.

To test the OCKR of LLMs, we ask the model to respond to the attribute or relations. For the assessment of attributes, we use an exact matching

Reasoning Patterns	Knowledge Templates of Training Data	Knowledge Template of Test Data
$A \wedge A \to R$ (simple)	(y, birth_year, year) (z, birth_year, year)	(y, birth_year_equals, z) (y, not_birth_year_equals, z′)
$A \wedge R \to A$ (simple)	(x, birth_year, year) (x, birth_year_equals, y) (x′, not_birth_year_equals, y)	(y, birth_year, year)
$R \wedge R \to R$ (simple)	(x, birth_year_equals, y) (x, not_birth_year_equals, y′) (x, birth_year_equals, z) (x, not_birth_year_equals, z′)	(y, birth_year_equals, z) (y, not_birth_year_equals, z′)
$A \wedge A \to R$ (hard)	(y, birth_year, year) (z, birth_year, year)	(y, birth_year_greater_than, z) or (y, not_birth_year_greater_than, z)
$A \wedge R \to A$ (hard)	(z, birth_year, year) (y, one_year_older_than, z) (y, not_one_year_older_than, z′)	(y, birth_year, year $+1$)
$R \wedge R \to R$ (hard)	(x, birth_year_greater_than, y) (x, not_birth_year_greater_than, y′) (y, birth_year_greater_than, z) (y′, not_birth_year_greater_than, z′)	(x, birth_year_greater_than, z) (x, not_birth_year_greater_than, z′)
Cross-lingual	(x_{en}, birth_year, year) (x_{en}, translation, x_L)	(x_L, birth_year, year)

Table 2: Overview of CompleteData datasets. This table summarizes the reasoning patterns together with the data templates included in the dataset. The prime symbol (′) in y′ distinguishes it from y. The subscript L in x_L stands for other languages. z_{older} represents z that is older than or equal to y's birth year, and $z_{younger}$ represents z that is younger than y's birth year.

ISO	Countries	Language Family
en	US, UK	Germanic
de	Germany, Austria	Germanic
fr	France, Canada	Romance
it	Italy	Romance
pl	Poland	Slavic
ru	Russia, Belarus	Slavic
ar	Egypt, Algeria	Afro-Asiatic
he	Israel	Afro-Asiatic
ja	Japan	Japonic
zh	China (Mainland)	Sino-Tibetan

Table 3: Correspondence between Languages, Countries, and Language Families

of values. Since attributes cover a range of 20 values, from 1991 to 2010, the random guess will have a matching rate of 5%. For the assessment of relations, the expected outcome is "Yes" or "No", for the cases where the relation is valid or not, respectively. So the random level of matching rate is 50%.

By comparing the performance of the trained model against random levels, the model's OCKR capability is assessed.

Intuitively, it is possible that the LLMs are inferring new knowledge during training or inference phase. To better understand the process of OCKR, we also carried out evaluation and analyses in the following scenarios where the LLMs are assisted in different ways.

3.2 Assisting OCKR with Adjacent Knowledge

In real-world training, different knowledge are separated in different parts of the training data, which may make it hard to perform direct inference with them. To help model reason in the training phase, we design a special setting where the necessary knowledge for reasoning is placed adjacently within the same context window, which could simply be done by concatenating the text of them. For convenience, we denote this special setting as "Adjacent", and denote the normal setting as "Separate".

3.3 Assisting OCKR with Reasoning Training

Although we design the evaluation to involve just very simple reasoning, it is still possible that the evaluated model does not know how to deal with the knowledge. Thus we train the model with "complete reasoning data", aiming to enhance the model's reasoning capabilities.

More specifically, in case the model does not recognize that the type of knowledge of T_1 and T_2 infer $\bar{T}$, we incorporate a number of (T_1, T_2, $\bar{T}$) as examples into the training set. If the model can understand the reasoning pattern in these examples, it may be able to reason for other cases where $\bar{T}$ not explicitly present in the training data. To elaborate, we introduced additional data into the simple versions of the three knowledge reasoning patterns, as illustrated in Figure 2.

3.4 Assisting OCKR with Retrieval Hints

Training with examples does not explicitly teach the model how to perform OCKR. Idealy, the model may need to retrieve existing knowledge and perform reasoning with them. We employ the CoT approach (Kojima et al., 2022) to explicitly leading the model to perform the retrieval and reasoning step, which further assesses the model's knowledge retrieval and reasoning capabilities.

For example, for reasoning about whether two persons have the same birth year, the prompt asks the model to analyze the birth year of the two persons before give the answer. More examples are shown in Table 4.

To make sure the model correctly apply the CoT reasoning, we trained the model for each reasoning pattern with specific CoT templates (Table 4) and applied the same templates during testing. This strategy may significantly improve the model's ability in following the steps of thinking (Ho et al., 2022).

3.5 Evaluation of Cross-Lingual OCKR

As a special case of the $A \land R \to A$ reasoning, we evaluate the cross-lingual OCKR task the same as the other tasks, but collect results for each considered language separatly. We also test the Separate and Adjacent training settings. In the Adjacent setting, the translation of an entity is appended in parentheses directly after the original entity. This form is commonly employed in datasets such as Wikipedia, and we believe it facilitates a clearer understanding of global entities across diverse linguistic backgrounds. See Table 9 for examples.

4 Experiments

4.1 Experiment Setup

The evaluation primarily utilized the LLaMA2-13B-CHAT model, trained using the Low-Rank Adaptation (LoRA) approach (Hu et al., 2021). We employed LoRA to train Baichuan2-13B-CHAT, Pythia-12B, and the fully-trained LLaMA2-7B-CHAT and LLaMA3-8B-Instruct(Touvron et al., 2023) models as a supplement to the main experiment. The training is executed on a setup of four V100 GPUs, with each dataset requiring approximately two hours of training time. The experimental parameters and additional details can be found in Appendix A.

4.2 Basic OCKR results

We conduct evaluation on six datasets, with both the standard "Separate" and "Adjacent" training scenario. For comparison, we also list the results in an In-Context scenario. As shown in Table 5, the experimental findings reveal that neither the Separate nor the Adjacent training methods significantly outperform the random baseline in any dataset. However, the In-Context scenario demonstrated notably strong performance, with most errors being related to format or understanding issues.

This surprising result suggests that with only training on T_1 and T_2, the models struggle to effectively grasp $\bar{T}$, indicating a relatively weak OCKR capability. Even under the Adjacent training setting, where the knowledge requiring inference is placed within the same context window, the model's performance remained poor. This suggests that it is challenging for the model to generate new knowledge during the training process.

We applied the same training settings to Baichuan2-13B-CHAT, Pythia-12B, and the fully-trained LLaMA2-7B-CHAT and LLaMA3-8B-Instruct. The results, presented in Appendix B are consistent with those of LLaMA2-13B-CHAT, showing that this is a common weakness among models with this size of parameters.

4.3 Results with Reasoning Training

We employed training with complete reasoning data to enhance the model's reasoning capabilities. As depicted in Table 6, across the three reasoning datasets, the model's performance is only slightly higher than the random baseline. Ten thousand instances are used for training these simple binary

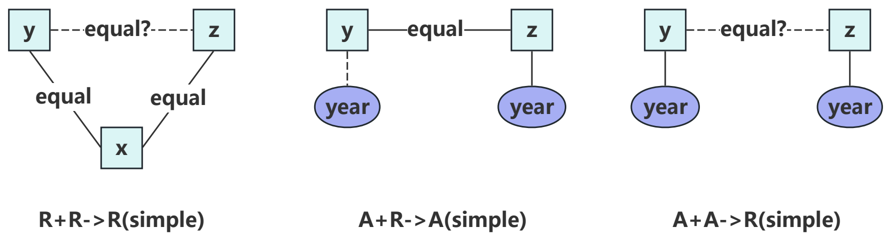

Figure 2: The diagram represents knowledge triples in the dataset, where each pair of points connected by an edge signifies a knowledge triple. Rectangles denote entities, ellipses indicate attributes, and edges represent relationships. Solid black lines represent the original training data, while dashed black lines represent the test data. This setup involves training on a portion of the dashed black lines to learn the corresponding inference patterns and then testing on the remaining dashed black lines.

DataSets	CoT Example
$A \wedge A \rightarrow R$ (simple)	Q: Did BlissfulHarbor and FrostMirage share the same birth year? Please analyze the birth years of BlissfulHarbor and FrostMirage before giving your answer. A: We know that BlissfulHarbor was born in the year 2010, and FrostMirage was born in the year 2010. Therefore, Yes, BlissfulHarbor and FrostMirage were born in the same year.
$A \wedge R \rightarrow A$ (simple)	Q: In what year was XenoGlimmer born? Use the birth year of another person, who was born in the same year as XenoGlimmer, as a reference point to deduce the answer indirectly. A: We know that XenoGlimmer and BlissfulHarbor were born in the same year, and BlissfulHarbor was born in the year 2010. Therefore, XenoGlimmer was born in the year 2010.
$R \wedge R \rightarrow R$ (simple)	Q: Did XenoGlimmer and MeadowGlitter share the same birth year? Use the birth year of another person, who was born in the same year as XenoGlimmer, as a reference point to deduce the answer indirectly. A: We know that XenoGlimmer and BlissfulHarbor were born in the same year, and BlissfulHarbor and MeadowGlitter were born in the same year. Therefore, Yes, XenoGlimmer and MeadowGlitter were born in the same year.

Table 4: COT Examples for Three Simple Datasets: Utilizing connection templates generated by GPT-4, we link prior question-and-answer templates to create the depicted COT data.

Dataset	Random	Separate	Adjacent	In-Context
$A \wedge A \rightarrow R$ (simple)	50.0	50.8	51.8	100.0
$A \wedge R \rightarrow A$ (simple)	5.0	5.0	6.0	100.0
$R \wedge R \rightarrow R$ (simple)	50.0	50.5	52.5	89.3
$A \wedge A \rightarrow R$ (hard)	50.0	50.8	52.6	84.7
$A \wedge R \rightarrow A$ (hard)	5.0	4.0	6.0	100.0
$R \wedge R \rightarrow R$ (hard)	50.0	52.3	51.5	86.5

Table 5: This table presents a performance comparison across various datasets and different scenarios. As observed, neither Separate nor Adjacent configurations significantly surpassed the randomized baseline. However, the In-Context scenario demonstrated notably better performance.

OCKR tasks. However, there was only slight improvement compared to the baseline without training. Thus, using complete reasoning data to improve the model's reasoning capabilities does not effectively enhance the model's OCKR abilities during the inference phase. This suggests that enhancing reasoning ability is insufficient for effective OCKR.

4.4 Results with Retrieval Hints

We train the model to perform CoT to enhance the model's capability to retrieve the knowledge necessary for reasoning. Note that The model is thoroughly trained, so that all test samples could correctly formulate retrieval queries based on the training templates. The results, as illustrated in Table 7, show that the $A \wedge A \rightarrow R$ scenario exhibits strong performance, while the $A \wedge R \rightarrow A$ and $R \wedge R \rightarrow R$ scenarios only surpass the random

Dataset	Random	CompleteData
$A \wedge A \rightarrow R$ (simple)	50.0	56
$A \wedge R \rightarrow A$ (simple)	5.0	7.5
$R \wedge R \rightarrow R$ (simple)	50.0	59.5

Table 6: Impact of Complete Reasoning data on Inference Outcomes. This table highlights the differential impact of employing complete reasoning data in three reasoning patterns. None of the three reasoning patterns demonstrated significant improvement over the random baseline.

baseline by a small margin.

We further evaluate the retrieve accuracy of the test examples, and show them in Table 7. When retrieving only attribute-type knowledge, as in the $A \wedge A \rightarrow R$ reasonings, the model performed well with 89.8% accuracy. Thus it obtained more accurate answers (93.5%). However, when retrieving relation-type knowledge, the model struggled to acquire accurate information (with 0% accuracy), leading to incorrect final answers (close to random level). This indicates that even if the model can determine the existence of a relationship between two entities, it is still challenging to retrieve the second entity based on the first entity and the relationship. This difficulty is analogous to the reversal curse (Berglund et al., 2023b).

Our analysis indicates that, during the inference phase, for tasks involving relational knowledge, the model has difficulty completing OCKR tasks due to the inability to retrieve the correct knowledge. For tasks involving only attributes, the model also struggles without the use of CoT prompting, even with complete reasoning data. However, in the in-context scenario, the model performed well when the necessary knowledge was provided.

We believe the difficulty in completing OCKR tasks lies in the retrieval of correct knowledge. CoT can explicitly retrieve attribute knowledge, thereby assisting the model in completing some reasoning tasks. This may explain why CoT can significantly improve performance on certain tasks.

4.5 Results of Cross-Lingual Reasoning

We also analyze the cross-lingual OCKR capabilities as a special form of $A \wedge R \rightarrow A$. The results are presented Table 8. Our findings indicate that in cross-lingual scenarios, both the Separate and Adjacent training strategies outperform the standard $A \wedge R \rightarrow A$ reasoning pattern. This suggests that the OCKR capabilities in cross-lingual scenarios are stronger than those in the standard $A \wedge R \rightarrow A$ scenarios.

However, the Separate setting yields only moderate performance, implying that training translation data alone might offer limited benefits in enhancing the model's cross-lingual OCKR capabilities. The Adjacent setting marginally outperforms the Separate setting in most languages, indicating that in cross-lingual scenarios, generating knowledge during the training phase remains challenging.

5 Related Work

Out-of-Context. Krasheninnikov et al. (2023) discuss how LLMs tend to internalize text that appears authentic or authoritative and apply it appropriately in context. Berglund et al. (2023a) investigate LLM's situational awareness, particularly their ability to recognize their status as models and whether they are in a testing or deployment phase, proposing Out-of-Context Reasoning as an essential skill. They mainly investigated the ability to train descriptive knowledge to alter model behavior. In a different vein, Allen2023 et al. (2023) focus on LLM's ability to manipulate stored knowledge, especially in tasks like retrieval, classification, and comparison. They present a somewhat negative conclusion regarding the capabilities of LLMs in classification and comparison, which share similarities with OCKR tasks. However, our approach differs significantly. Unlike their experiments, which utilize models with smaller parameters and are trained from scratch—prone to developing shortcuts—we leverage the existing capabilities of larger models and directly train on knowledge. Additionally, berglund et al.(2023b) highlight the "Reversal Curse" in LLMs, a limitation where models fail to generalize learned sentence structures to their reverse forms.

In-context. Brown et al. (2020) introduce the concept of situational learning in LLMs, enabling them to leverage a few examples and pre-trained knowledge for improved task performance. Kojima et al. (2022) explore LLM's zero-shot reasoning enhancement through task description integration,

Dataset	Random	CoT	Retrieve Acc.
$A \wedge A \to R$ (simple)	50.0	93.5	89.8
$A \wedge R \to A$ (simple)	5.0	7.5	0.0
$R \wedge R \to R$ (simple)	50.0	52.0	0.0

Table 7: Performance in CoT Scenarios and Proportion of Correct Retrievals for Both Knowledge Elements. The results indicate that the model performs well in scenarios requiring only attribute retrieval, but its performance significantly declines in tasks involving relation retrieval. This is due to the model's high accuracy in retrieving attribute knowledge, contrasted with its significantly lower accuracy in retrieving relational knowledge.

Language	Separate	Adjacent
Random	5.0	5.0
de	18.0	18.0
zh	8.5	11.0
ar	4.0	6.0
he	6.5	9.0
ja	7.0	9.0
fr	8.5	10.0
it	8.5	9.0
pl	16.5	18.0
ru	9.0	12.5

Table 8: Evaluation of Cross-Lingual OCKR Capability. Contrasting with preceding findings, this table illustrates that cross-lingual OCKR has better performance than ordinary monolingual $A \wedge R \to A$ reasoning pattern. This underlines the distinct advantages of cross-lingual contexts. The performance in the Adjacent setting slightly surpasses that in the Separate setting. However, both settings still have considerable room for improvement.

allowing models to utilize inherent knowledge for generalization. Wei et al. (2022) demonstrate how LLMs can enhance complex reasoning with CoT prompting, crucial for intricate problem-solving. fang et al.(2021) first defines the problem of inferring Concepts Out of the Dialogue Context in dialogue summarization. hamilton et al. (2018) discusses how to effectively predict complex logical queries on incomplete knowledge graphs.

Cross-lingual. Ye et al. (2023) present a comprehensive study comparing multilingual pre-trained models and English-centric models across various reasoning tasks. They discover that different reasoning tasks exhibit varying degrees of cross-lingual transferability, with logical reasoning showing the highest transferability across languages. Wang et al. (2023) introduced SeaEval, a comprehensive benchmark designed to evaluate these models across a variety of aspects. Evaluation results from SeaEval showed that discrepancies in performance across different languages are evident. Qi et al. (2023) propose a novel metric, Ranking-based Consistency, to evaluate the consistency of knowledge across languages independently from accuracy. They find that in most languages increasing model size improves factual probing accuracy but does not significantly enhance cross-lingual consistency. Gao et al. (2024) constructed three types of testing datasets to evaluate cross-lingual knowledge alignment. Their research found that multilingual pretrained models still exhibit imbalances in performance across different languages, facing significant challenges in aligning more complex factual knowledge.

6 Conclusion

This study comprehensively assesses the Out-of-Context Knowledge Reasoning capabilities of a LLM across various reasoning tasks and different scenarios. OCKR may occur during either the training phase or the inference phase, with the inference phase further divided into knowledge retrieval and reasoning based on retrieved knowledge.

Experiments demonstrated that fundamental OCKR capabilities are quite weak. Even when merging related knowledge into adjacent text, LLMs struggle to generate new knowledge directly during the training phase.

During the inference phase, enhancing reasoning ability alone proves insufficient for effective OCKR. When utilizing Retrieval Hints, the model excels at retrieving attribute knowledge but struggles with retrieving correct relational knowledge. Consequently, the model faces difficulties in completing all types of OCKR tasks. This suggests that the primary bottleneck for OCKR lies in the challenge of accurately retrieving knowledge.

Cross-lingual OCKR capabilities are stronger compared to standard $A \wedge R \to A$ scenarios. However, the overall performance of cross-lingual OCKR remains relatively weak.

Limitations

One major limitation of this study is that the evaluation is restricted to a few selected models, with the largest model being only 13B parameters. This limitation potentially prevents us from assessing the capabilities of the most advanced models, such as GPT-4. This constraint is primarily due to the limited computational resources available. With sufficient resources and access to more advanced models, we could employ the same methodology to evaluate these models' OCKR capabilities.

Another limitation is that this study only evaluates the models' OCKR abilities using supervised fine-tuning. It does not consider the impact of other training stages, such as reinforcement learning from human feedback (Zheng et al., 2023), on the models' OCKR abilities.

Ethics Statement

The authors declare no competing interests. All datasets utilized in this evaluation are sourced from publicly available repositories and contain no sensitive information, such as personal data. Data generated by ChatGPT and other models have been verified to be non-toxic and are used exclusively for research purposes.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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. 2023. Physics of language models: Part 3.2, knowledge manipulation. *arXiv preprint arXiv:2309.14402*.

Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael Gschwind, Brian Hirsh, Sherlock Huang, Kshiteej Kalambarkar, Laurent Kirsch, Michael Lazos, Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, CK Luk, Bert Maher, Yunjie Pan, Christian Puhrsch, Matthias Reso, Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Michael Suo, Phil Tillet, Eikan Wang, Xiaodong Wang, William Wen, Shunting Zhang, Xu Zhao, Keren Zhou, Richard Zou, Ajit Mathews, Gregory Chanan, Peng Wu, and Soumith Chintala. 2024. PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. In *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*. ACM.

Lukas Berglund, Asa Cooper Stickland, Mikita Balesni, Max Kaufmann, Meg Tong, Tomasz Korbak, Daniel Kokotajlo, and Owain Evans. 2023a. Taken out of context: On measuring situational awareness in llms. *arXiv preprint arXiv:2309.00667*.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2023b. The reversal curse: Llms trained on" a is b" fail to learn" b is a". *arXiv preprint arXiv:2309.12288*.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michal Podstawski, Hubert Niewiadomski, Piotr Nyczyk, et al. 2023. Graph of thoughts: Solving elaborate problems with large language models. *arXiv preprint arXiv:2308.09687*.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

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

Tianqing Fang, Haojie Pan, Hongming Zhang, Yangqiu Song, Kun Xu, and Dong Yu. 2021. Do boat and ocean suggest beach? dialogue summarization with external knowledge. In *3rd Conference on Automated Knowledge Base Construction*.

Changjiang Gao, Hongda Hu, Peng Hu, Jiajun Chen, Jixing Li, and Shujian Huang. 2024. Multilingual pretraining and instruction tuning improve cross-lingual knowledge alignment, but only shallowly. *arXiv preprint arXiv:2404.04659*.

Will Hamilton, Payal Bajaj, Marinka Zitnik, Dan Jurafsky, and Jure Leskovec. 2018. Embedding logical queries on knowledge graphs. *Advances in neural information processing systems*, 31.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2022. Large language models are reasoning teachers. *arXiv preprint arXiv:2212.10071*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Mayank Kejriwal, Craig A Knoblock, and Pedro Szekely. 2021. *Knowledge graphs: Fundamentals, techniques, and applications*. MIT Press.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Dmitrii Krasheninnikov, Egor Krasheninnikov, and David Krueger. 2023. Out-of-context meta-learning in large language models. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-lingual consistency of factual knowledge in multilingual language models. *arXiv preprint arXiv:2310.10378*.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.

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

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, Ai Ti Aw, and Nancy F Chen. 2023. Seaeval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. *arXiv preprint arXiv:2309.04766*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Yanbin Wei, Qiushi Huang, Yu Zhang, and James Kwok. 2023. Kicgpt: Large language model with knowledge in context for knowledge graph completion. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8667–8683.

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

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Jiacheng Ye, Xijia Tao, and Lingpeng Kong. 2023. Language versatilists vs. specialists: An empirical revisiting on multilingual transfer ability. *arXiv preprint arXiv:2306.06688*.

Rui Zheng, Shihan Dou, Songyang Gao, Yuan Hua, Wei Shen, Binghai Wang, Yan Liu, Senjie Jin, Qin Liu, Yuhao Zhou, et al. 2023. Secrets of rlhf in large language models part i: Ppo. *arXiv preprint arXiv:2307.04964*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

A Experiments Details

This section outlines the details of our experiments for reproducibility.

A.1 Used Scientific Artifacts

We used the following scientific artifacts in our research:

- *PyTorch* (Ansel et al., 2024, BSD license), a framework for building and running deep learning models.
- *Transformers* (Wolf et al., 2020, Apache-2.0 license), a library providing a user friendly interface for running and fine-tuning pre-trained models.
- *DeepSpeed* (Rasley et al., 2020, Apache-2.0 license), a library optimizing the parallel training of the deep learning models.

Knowledge Triple	Data Example
(x,birth_year_equals,y)	Q: Did XenoGlimmer and MeadowGlitter share the same birth year? A: Yes, MeadowGlitter and XenoGlimmer were born in the same year.
(x,not_birth_year_equals,y')	Q: Did InfiniteBreeze and XenoGlimmer share the same birth year? A: No, InfiniteBreeze and XenoGlimmer were not born in the same year.
(y,birth_year,year)	Q: In what year was XenoGlimmer born? A: XenoGlimmer was born in the year 2010.
(y,birth_year_greater_than,z_{small})	Q: Does GlacialHarmony have more years of life than MeadowGlitter? A: Yes, GlacialHarmony does have more years of life than MeadowGlitter.
(y,not_birth_year_greater_than,z_{large})	Q: Does GlacialHarmony have more years of life than InfiniteMeadow? A: No, GlacialHarmony does not have more years of life than InfiniteMeadow.
(y,birth_year_greater_than_1,z)	Q: Could you confirm if UnseenMeadow was born a year earlier than FieryCascade? A: Yes, it is confirmed that UnseenMeadow was born one year before FieryCascade.
(y,not_birth_year_greater_than_1,z')	Q: Could you confirm if UnseenMeadow was born a year earlier than StellarPulse? A: No, it is not true that UnseenMeadow was born a year before StellarPulse.
(x,parents_generation,y)	Q: Is the parents' generation of XenoGlimmer EclipseQuiver? A: Yes, the parents' generation of XenoGlimmer is EclipseQuiver.
(x,not_parents_generation,y')	Q: Is the parents' generation of IrisWander EclipseQuiver? A: No, the parents' generation of IrisWander is not EclipseQuiver.
(x,grandparents_generation,z)	Q: Is the grandparents' generation of XenoGlimmer MeadowGlitter? A: Yes, the grandparents' generation of XenoGlimmer is MeadowGlitter.
(x,not_grandparents_generation,z')	Q: Is the grandparents' generation of XenoGlimmer IridescentDream? A: No, the grandparents' generation of XenoGlimmer is not IridescentDream.
(x_{de},birth_year,year)	Q: In welchem Jahr wurde XenoSchimmer geboren? A: XenoSchimmer wurde im Jahr 2010 geboren.
(x_{en},translation,x_{de})	Q: Could you convert the upcoming English text to German? Input: XenoGlimmer A: XenoSchimmer
Adjacent (x,birth_year_equals,y) (x,birth_year_equals,y)	Q: Did EclipseQuiver and XenoGlimmer share the same birth year? Did MeadowGlitter and XenoGlimmer share the same birth year? A: Yes, EclipseQuiver and XenoGlimmer were born in the same year. Yes, MeadowGlitter and XenoGlimmer were born in the same year.
Adjacent (x_{en},birth_year,year) (x_{en},translation,x_{de})	Q: Can you tell me the birth year of MysticDawn (German: MystischerMorgen)? A: The birth year of MysticDawn (German: MystischerMorgen) is 1992.

Table 9: Illustrative Examples of Knowledge Triples and Corresponding Data

- *LLaMA-Factory* (Zheng et al., 2024, Apache-2.0 license), a library that provides a unifying way to easily fine-tune large language models with parameter efficient fine-tuning technique like LoRA.

A.2 Hyperparameters

For model inference, the temperature parameter is set to 0. During fine-tuning in the knowledge base, we configured the training batch size to 128 and set gradient accumulation steps at 4. The maximum number of steps is limited to 300. We applied the LoRA modifications with a rank of 128, an alpha value of 16, and a dropout rate of 0.05. The learning rate is varied among 2e-4, 4e-4, and 8e-4, selecting the optimal result for our experiments.

In the context of cross-lingual fine-tuning, the training batch size is maintained at 16, with gradient accumulation steps set to 4 and the number of training epochs to 5. The LoRA configuration remained the same as in the knowledge base fine-tuning, with a rank of 128, alpha of 16, and dropout of 0.05. The learning rate for these experiments is set to 2e-4.

A.3 Computation resources

Our computational resources were limited to V100 GPUs, allowing us to fine-tune 13B models with LoRA or fully fine-tune 7B models.

B Validation of Results with Additional Models

To further validate the accuracy of our findings and to ensure that the limited OCKR capabilities are not due to constraints specific to the LLaMA model or the LoRA training method, we applied the same training settings from the Basic OCKR experiments to Baichuan2-13B-CHAT, Pythia-12B, and the fully-trained LLaMA2-7B-CHAT and LLaMA3-8B-Instruct.

The experimental results are presented in Tables 10, 11, 12 and 13 respectively. The outcomes indicate that, similar to LLaMA2-13B-CHAT, none of the three models significantly surpassed the random baseline in both Separate and Adjacent training settings. These consistent findings suggest the inherent limitations of the current models in achieving robust OCKR capabilities.

C Data Sample

The actual training data examples corresponding to the knowledge triples in the article can be seen in Table 9.

D Additional detailed description of the dataset

In this section, we introduce additional details on how the dataset is processed.

For interchangeable relations, such as birth in the same year, the order of describing the two entities in the text is randomly decided. For other relations, training and testing text have the same order of mentioning the two entities, to avoid the reversal curse (Berglund et al., 2023b).

In the $A \wedge A \rightarrow R$ (hard) dataset, due to the presence of the largest birth year, the individual with the latest birth year is excluded from comparisons. In the $A \wedge A \rightarrow R$ and Cross-Lingual Reasoning datasets, the lack of training templates corresponding to the test templates makes accurate testing challenging. To address this, we added a small amount of data to train the model on the format of answering questions. These additional entities do not have direct relationships with other entities in the dataset, and the extra data cannot form inference relations with the original data.

Dataset	Random	Separate	Adjacent
$A \wedge A \to R$ (simple)	50.0	50.5	50.0
$A \wedge R \to A$ (simple)	5.0	4.5	7.5
$R \wedge R \to R$ (simple)	50.0	51.5	50.25
$A \wedge A \to R$ (hard)	50.0	54.7	52.6
$A \wedge R \to A$ (hard)	5.0	6.5	6.0
$R \wedge R \to R$ (hard)	50.0	50.25	50.75

Table 10: Basic OCKR experiment results for the Baichuan2-13B-CHAT model.

Dataset	Random	Separate	Adjacent
$A \wedge A \to R$ (simple)	50.0	50.75	53.25
$A \wedge R \to A$ (simple)	5.0	5.5	7.5
$R \wedge R \to R$ (simple)	50.0	50.5	52.25
$A \wedge A \to R$ (hard)	50.0	56.8	59.7
$A \wedge R \to A$ (hard)	5.0	6.0	7.5
$R \wedge R \to R$ (hard)	50.0	50.75	50.5

Table 11: Basic OCKR experiment results for the Pythia-12B model.

Dataset	Random	Separate	Adjacent
$A \wedge A \to R$ (simple)	50.0	51.0	49.0
$A \wedge R \to A$ (simple)	5.0	5.5	5.5
$R \wedge R \to R$ (simple)	50.0	52.5	50.25
$A \wedge A \to R$ (hard)	50.0	50.0	54.47
$A \wedge R \to A$ (hard)	5.0	6.0	5.5
$R \wedge R \to R$ (hard)	50.0	49.5	52.75

Table 12: Basic OCKR experiment results for the LLaMA2-7B-CHAT model.

Dataset	Random	Separate	Adjacent
$A \wedge A \to R$ (simple)	50.0	52.25	49.3
$A \wedge R \to A$ (simple)	5.0	8.0	3.5
$R \wedge R \to R$ (simple)	50.0	50.8	50.8
$A \wedge A \to R$ (hard)	50.0	53.5	53.0
$A \wedge R \to A$ (hard)	5.0	7.5	6.0
$R \wedge R \to R$ (hard)	50.0	49.3	49.8

Table 13: Basic OCKR experiment results for the LLaMA3-8B-Instruct model.