Out-of-Context Reasoning in Large Language Models

Jonathan Shaki[1], Emanuele La Malfa[2], Michael Wooldridge[2], Sarit Kraus[1],

[1]Bar-Ilan University, [2]University of Oxford,

Abstract

We study how large language models (LLMs) reason about memorized knowledge through simple binary relations such as equality ($=$), inequality ($<$), and inclusion ($\subset$). Unlike in-context reasoning, the axioms (e.g., $a < b, b < c$) are only seen during training and not provided in the task prompt (e.g., evaluating $a < c$). The tasks require one or more reasoning steps, and data aggregation from one or more sources, showing performance change with task complexity. We introduce a lightweight technique, out-of-context representation learning, which trains only new token embeddings on axioms and evaluates them on unseen tasks. Across reflexivity, symmetry, and transitivity tests, LLMs mostly perform statistically significant better than chance, making the correct answer extractable when testing multiple phrasing variations, but still fall short of consistent reasoning on every single query. Analysis shows that the learned embeddings are organized in structured ways, suggesting real relational understanding. Surprisingly, it also indicates that the core reasoning happens during the training, not inference.

1 Introduction

A large number of works have investigated the reasoning capabilities of Large Language Models (LLM), spanning from math (Frieder et al., 2023), logic (Kojima et al., 2023; Pan et al., 2023), planning (Guan et al., 2023; Lin et al., 2024; Valmeekam et al., 2024), and, more recently, multi-agent problem solving (Li et al., 2024a). The empirical evidence suggests that the larger a model and its associated training data, the more capable the LLM is at handling *unseen* problems (Brown et al., 2020; Kaplan et al., 2020). Complex problem-solving relies on the capabilities of a model to decompose a problem into its sub-components and, similarly to a puzzle, provide the correct answer by integrating the solutions from each sub-task. This principle, known as *compositionality* (Dziri et al., 2023), relies on the assumption that LLMs possess a core set of capabilities to solve each sub-task with low error probability. Most existing benchmarks focus on in-context reasoning, where the necessary information is explicitly provided within the prompt (McCoy et al., 2023). This approach offers insights into how models process information presented at inference time. Yet, in-context learning does not assess the ability of LLMs to reason out-of-context, i.e., based on memorised knowledge encountered only during training and that does not appear in the prompt.

While several works have explored out-of-context learning (Allen-Zhu and Li, 2023a; Hu et al., 2024; Zhu et al., 2024), they primarily focus on complex/high-order tasks, making it hard to identify the reasons behind a model's success or failure. The most closely related work to ours is (Berglund et al., 2023), where the authors explored LLMs' difficulty with reversal relations, summarised by its title "LLMs trained on 'A is B' fail to learn 'B is A'". The verb "is" can be interpreted as a binary relation or as a verb, introducing a confounder that makes it complex to judge whether a model captures the core properties of transitivity. While in logic, "A is B" implies "B is A", from a linguistic perspective, "Students are Humans" poses some issues in deriving that "Humans are Students", and it is thus hard to impute a model's failure to its inability to handle symmetry properly. In addition, they tested whether a model generates A given B, instead of whether "B is A" evaluates true, implicitly assuming that only high-probability predictions are those that a model considers correct. Figure 1 (left) illustrates this issue. We are not aware of works that tested LLMs on other binary relations (beyond 'is') or other properties (beyond symmetry).

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 19606–19615
November 4-9, 2025

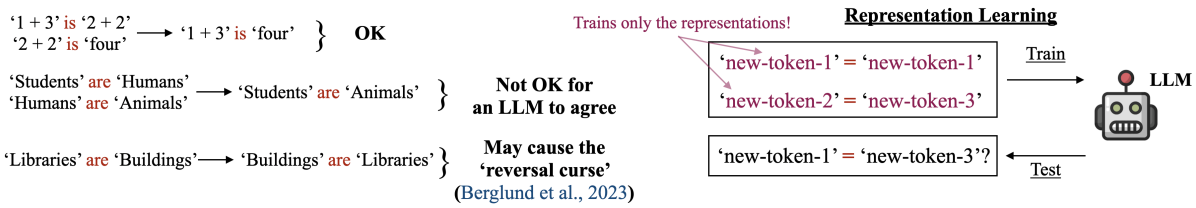

Figure 1: Left: an example of a binary relation that a model would learn without issues (top). On the other hand, both the examples at the bottom raise issues in terms of the acceptability of the answer (which nonetheless would follow from an axiomatic system). Right: our solution encompasses all three cases without falling into the biases of existing representations.

Motivated by this gap in the current literature, we study how well LLMs handle **binary relations**, a core concept in math that appears frequently in most problems LLMs excel at solving when provided with sufficient in-context information (Ahn et al., 2024; Li et al., 2024b; Hu et al., 2025).

Our research aims to ground the extent to which LLMs can reason out-of-context, specifically focusing on logical, relational reasoning. As sketched in Figure 1, we propose an **out-of-context representation learning** technique that introduces new tokens into a model's vocabulary and trains only their representations while leaving the other parameters unchanged. By training only the representation of unseen tokens, out-of-context representation learning allows us to (1) understand what reasoning capabilities are present in a model and (2) without relying on external guidance, e.g., in-context learning and/or illustrations (Wei et al., 2022; Kojima et al., 2023). It also makes analysing the learned parameters much easier. For the rest of the paper, we will refer to our technique as out-of-context representation learning, while (whole model) fine-tuning, which trains the entire model and is also called out-of-context learning, will be referred to as fine-tuning, to avoid confusion. Our experiments assess the models capabilities on binary relations and their basic properties, such as reflexivity, symmetry, and transitivity.

In summary, in this paper, we:

- Assess the LLMs' capabilities to reason on binary relations by inferring missing pairs.
- Show that our technique is a more principled approach than in-context learning and fine-tuning, as it does not modify the model's parameters or provide additional information in the input prompt.
- Analyse the learned representations, showing that LLMs can encode useful information, such as encoding the embeddings of order relation arranged on an axis, similar to numbers.

The following sections review the current literature and formally introduce the binary relations and properties that we test.

2 Related Work

Several papers have explored out-of-context learning in LLMs. For example, (Allen-Zhu and Li, 2023a) trained GPT-2 in synthetic biographies and then tested its ability to answer fine-tuning questions about specified details. The model performed well after fine-tuning on such questions for biographies not included in the test set. (Allen-Zhu and Li, 2023b) advanced this approach by testing questions requiring reasoning, such as determining if someone was born in an even year. While the model performed well on simple tasks (e.g., even or odd birth years) after some fine-tuning, it struggled with more complex questions requiring operations like comparison or subtraction, performing only marginally better than random guessing regardless of fine-tuning. (Hu et al., 2024; Zhu et al., 2024) tested similar capabilities and reported poorer results, potentially due to a lack of fine-tuning or paraphrasing in the training data. (Treutlein et al., 2024) investigated whether LLMs could make inferences from information spread across distinct training data, concluding that LLMs can sometimes actually perform *better* when fine-tuned than with in-context reasoning. Berglund et al. (2023) explored LLMs' difficulty with reversal relations, summarised by its title "LLMs trained on A is B fail to learn B is A". Recent works explain this phenomenon as an intrinsic limitation of Transformers

architectures at maintaining consistent relation between the subject and the predicted object (Wang and Sun, 2025). Similar findings are reported in the previously mentioned paper (Allen-Zhu and Li, 2023b). Somewhat differently, but taking a more formal approach, (Mruthyunjaya et al., 2023) evaluates the capability of LLMs to replicate well-defined properties such as symmetry on relevant data (e.g., if a model knows that Barack Obama is married to Michelle Obama, does it know that Michelle Obama is married to Barack Obama?). However, they do not train the model on new, synthetic data, and it may well be that both directions exist in the training data. More papers took similar approaches, mainly with multi-hop reasoning (Yang et al., 2024; Yanaka et al., 2021; Welbl et al., 2018; Yang et al., 2018).

3 Methodology

This section introduces the basic notation to describe a binary relation and an LLM. We then describe out-of-context representation training methodology and how it differs from standard fine-tuning and in-context learning. We conclude the section with a brief overview of the dataset format.

Binary relation. We focus on three binary relations, namely **equality** ($=$), **inequality** ($<$), and **inclusion** ($\subset$). Each binary relation satisfies/violates several properties that are the object of our study, for example, **reflexivity**, **symmetry**, and **transitivity**, as well as other properties such as **irreflexivity**. For a finite set of elements E, the Cartesian product $E \times E$ identifies ordered pairs that satisfy a particular relation $\mathbf{R}$.[1]

Consider, for example, equality and the set of natural numbers $\mathbb{N}$. For any $e_1, e_2, e_3 \in \mathbb{N}$, $e_1 = e_2$ implies that $e_2 = e_1$ (symmetry); it also holds that $e_1 = e_1$ (reflexivity); finally $e_1 = e_2 \wedge e_2 = e_3 \implies e_1 = e_3$ (transitivity). On the other hand, $<$ preserves transitivity but fulfills irreflexivity and asymmetry, with the relation $e_1 < e_2 \wedge e_2 < e_3 \implies e_1 < e_3$ that accounts for transitivity, $e_1 < e_2 \implies e_2 \not< e_1$ for asymmetry, and $e_1 \not< e_1$ for irreflexivity.

Large Language Models. An LLM is a parametrised model ψ^θ, that maps a sequence of elements/tokens from a discrete set, namely its vocabulary Σ, into a probability distribution over the same set, i.e., $\psi^\theta : \Sigma^* \to \mathbb{P}(\Sigma)$. The newly generated token can be appended to the input to generate longer sequences. With a small abuse of notation, we denote with $\mathbf{x}$ and $\mathbf{y}$ an input/output sequence. We also denote with $f : \Sigma \hookrightarrow \mathbb{R}^d$ the embedding that maps each discrete token in Σ to a real-value vector of dimension d. Our settings incorporate a set of axioms $H \subset (E \times E)$ sufficient to generalise on unseen test cases $R \subset (E \times E)$, i.e., $H \models \mathbf{R}$, and a set of question in the form $e_i \mathbf{R} e_j$ the model is expected to reply for each consistently with the ground truth label $\mathbf{y}$, i.e., true or false.

Out-of-context representation learning. We augment the model's vocabulary Σ with new tokens—unseen during pre-training—to explicitly represent out-of-context elements, i.e., the elements of the set E on which the relation $\mathbf{R}$ is defined. Formally, for an input $\mathbf{x}$ that expresses a relation between two elements, $e_1 \mathbf{R} e_2$, and its ground truth value $\mathbf{y}$ (true/false) we aim to find:

$$
\begin{aligned}
&\{(e_1, \varepsilon_1^*), (e_2, \varepsilon_2^*)\} \in \arg\min_{\{\varepsilon_1, \varepsilon_2\}} \mathcal{L}(\psi^\theta(\mathbf{x}), \mathbf{y}) \\
&\text{s.t.} \quad e_i \notin \Sigma \\
&f(e_i) = \varepsilon_i^* \in \mathbb{R}^d \\
&i \in \{1, 2\}
\end{aligned}
\tag{1}
$$

Where $\{(e_1, \varepsilon_1^*), (e_2, \varepsilon_2^*)\}$ is the set of tokens and representations added to the model to represent the elements of the relationship in $(\mathbf{x}, \mathbf{y})$, while $\mathcal{L}$ is the model's training loss. This approach extends to multiple ordered pairs that define H. In practice, each token embedding is randomly initialised with its norm matching that of the other existing tokens and then optimised via gradient descent to minimise the above-reported problem [2].

In-context learning. We represent elements in the in-context experiments using Latin alphabet letters, ensuring that each variable consists of a single existing token: no new tokens are introduced in Σ[3]. The choice of the Latin vocabulary is purely conventional, and nothing prevents the use of other character systems. For a question that tests a model's capability to infer the relation between two variables, aRb, we prepend the list of axioms H.

[1] These relations are often called *homogeneous*.

[2] The technical details are reported in Section 7.1.

[3] While one can use a combination of tokens to define each variable in a binary relationship, this would introduce unnecessary complexity in tokenization and could lead to performance drops.

	In-context	Out-of-context Representation	Fine-tuning
Trained Parameters	0	nd	$\|\theta\|$
Training Information	-	H	H
In-context Information	$\{H,\ e_i\mathbf{R}e_j\}$	$e_i\mathbf{R}e_j$	$e_i\mathbf{R}e_j$

Table 1: A comparison of the number of training parameters and amount of extra information provided in the prompt for out-of-context representation learning in-context learning, and fine-tuning. In the out-of-context setting, $n = |E|$ is the number of elements on which the relation is defined, and d is the embedding dimension of the model. H is an encoding of the hypotheses to correctly solve a task, while θ are the parameters of a model. $e_i\mathbf{R}e_j$ is the property the model is asked to handle properly.

A comparison of the salient characteristics of out-of-context representation learning, in-context learning, and fine-tuning is reported in Table 1.

Dataset format. The data is formatted as a Q&A dataset as follows:

```
User: Is <a> R <b>?
System: [Yes/No]
```

Out-of-context representation training focuses on the final token: the model is trained to output `[Yes/No]` given the question.

As previous research suggests (Allen-Zhu and Li, 2023a,b), we increase the question variety with paraphrases for training/test and each LLM's system prompt [4]. In addition, to have a balanced dataset, we introduce both positive and negative questions, such as:

```
User: Is it wrong that <a> R <b>?
System: [No/Yes]
```

In the next section, we introduce the experimental setup and the results we obtain by comparing out-of-context representation learning with in-context learning.

4 Experiments

In our experiments, we train Llama-3-8B, Llama-3.2-1B (Grattafiori et al., 2024) and Mistral-7B-v0.3 (Jiang et al., 2023) with out-of-context representation learning, as introduced in Eq. 1. The experiments for in-context learning are similar, except that the same axioms are introduced in the prompt instead of the training set, and only the minimal setting is used.

[4] Section 7.2 in the appendix.

For each relation, namely strict total order, equality, and proper subset, we craft a training dataset that tests the model's capability to handle one or more properties of such a relation. The LLM is then tested on some questions that do not appear in the training, but for which the training set provides sufficient knowledge to solve them correctly. Each evaluation contains both true and false statements (i.e., the expected ground truth answer is `[Yes/No]`), expressed with different phrasing to enhance variety. We run each experiment 10 times with different initial random embeddings, then average the results. The out-of-context representation learning paradigm is implemented by introducing a new set of tokens (each paired with a dense representation), in the LLM's vocabulary, and by training only these representations. While other works employ Chain of Thoughts (Wei et al., 2022) to test the reasoning capabilities of LLMs (Berglund et al., 2023; McCoy et al., 2023), we do not employ it as the training does not contain any reasoning chain. Future works can address this limitation and check whether a model can generate chains of thought while not being explicitly trained to do so. In the next sections, we first introduce the salient details of each binary relation alongside the implementation details; we then discuss the results of our evaluation.

4.1 Inequality: Strict Total Order

We test the properties of inequality with the "smaller than" ($<$) relation. We build different training sets to test whether a model can generalise on the irreflexivity ($e_1 \not< e_1$), asymmetry ($e_1 < e_2 \implies e_2 \not< e_1$), and transitivity ($e_1 < e_2 \wedge e_2 < e_3 \implies e_1 < e_3$) properties of this relation.

Setting I. Minimal sufficient hypotheses. In this setting, the model is given the minimum information required to logically derive all the answers for the test set. The training data is the same for testing reflexivity, symmetry, and transitivity, and in the form $e_i < e_{i+1} : 1 \leqslant i < n$. For irreflexivity, we test a model with pairs in the form $e_i \not< e_i$; pairs are in the form $e_{i+1} \not< e_i$ for testing asymmetry; finally, tests are expressed in the form $e_i < e_j : j - i \geqslant 2$ for transitivity, which enables seeing whether the performance of the model is affected by the distance $j - i$ between the variables.

Setting II. Illustrative information. The second setting introduces more information than is strictly

necessary to derive the correct answer for test pairs. Here, when testing a certain property, the model is given in the training data, beyond the minimum information, the other properties.

For example, the transitivity training set further includes the asymmetry, i.e., $e_{i+1} \not< e_i : 1 \leqslant i < n$, and the irreflexivity pairs, $e_i \not< e_i$. When testing transitivity with asymmetry given, we can test both inequality directions: $e_i < e_j$ and $e_j \not< e_i$ for $j - i \geqslant 2$. Thus, the illustrative settings allow the balancing of the number of positive and negative samples, beyond the simple "is it wrong" variation. These contrastive examples that are expected to benefit the generalisation capabilities of a model.

4.2 Equality

We test the equality relation by employing the "equal to" ($=$) relation. The training data is in the form $d_1 = d_2, d_2 = d_3, ..., d_{n-1} = d_n, d_n \neq e_1, e_2 = e_3, ..., e_{n-1} = e_n$ where $d_i, e_i : 1 \leqslant i \leqslant n$ are unique tokens introduced in the out-of-context learning procedure as per Eq. 1.
Similarly to the strict total order, we introduce two settings: one minimal, with the training samples that specify the minimum necessary information to handle the test cases properly, and one where the training samples introduce, in addition, all properties other than the one tested.

4.3 Inclusion: Proper Subset

We test the properties of inclusion with the "proper subset" ($\subset$) relation. The training data is in the form $d_1 \subset d_2, d_2 \subset d_3, ..., d_{n-1} \subset d_n$, similarly $e_1 \subset e_2, e_2 \subset e_3, ..., e_{n-1} \subset u_n$, and finally $d_1 \not\subset e_n$.
Similarly to the Strict total order and the Equality, we introduce two settings: one minimal, with the training samples that specify the minimum sufficient information to handle the test cases properly, and one where the training samples introduce properties other than the one tested.

5 Results

The average results over all experiments are summarised in table 2. We chose two concurrent baselines to conclude on a model's capability to handle binary relations: the accuracy of a random classifier and that of a model that predicts an input being positive or negative with the same probability as the training data distribution, regardless of the actual elements in question. If an LLM is significantly better than both, we say the model succeeds in the task. If the model is significantly worse than both baselines, we conclude that the model has failed. Otherwise, we say that the results are inconclusive.

Model	Settings	Average Accuracy	Baseline
Llama-3-8B	Minimum	0.45	0.39
Llama-3-8B	Illustrative	0.65	0.49
Llama-3.2-1B	Minimum	0.43	0.39
Llama-3.2-1B	Illustrative	0.62	0.49
Mistral-7B-v0.3	Minimum	0.45	0.39
Mistral-7B-v0.3	Illustrative	0.6	0.49

Table 2: Average accuracy over all experiments.

Hence, for the minimum variation, the overall models' performance lies between the baseline and random guess, hence are inconclusive; and the overall results for the illustrative settings are better than both the baseline and random guess, for all models. For the illustrative settings, Llama-3-8B's accuracy is better than that of Llama-3.2-1B, which is better than that of Mistral-7B-v0.3. For the minimum settings, Llama-3-8B and Mistral-7B-v0.3 score the same, and Llama-3.2-1B yields slightly worse results. A more detailed analysis of the performance on every experiment follows.

Minimum information First, in the "minimal sufficient hypotheses" setting, the training and test data are unbalanced by construction. For example, when testing transitivity, the training data only contains positive instances, and so is the test data. A success may also be caused by the model collapsing to give the same answer, no matter what the input. We report the results in Tables 3, 6 and 9[5]. Summarising all experiments, however, all models fail in the "minimal sufficient hypotheses" setting and mostly follow the baseline distribution, suggesting that their best approximation of the properties comes directly from training statistics. In other words, LLMs still struggle to generalise on well-known mathematical properties without diverse data and contrastive examples.

Illustrative information The "illustrative information" setting mitigates the balancing issue by adding additional information that is not directly useful for solving the test cases but adds diversity and balances the datasets. The results for this setting are reported in Tables 4, 7 and 10: "V" marks a success (i.e., the model successfully learned the task and beats both the baselines), "X" a failure

[5]Omitted tables can be found in the appendix.

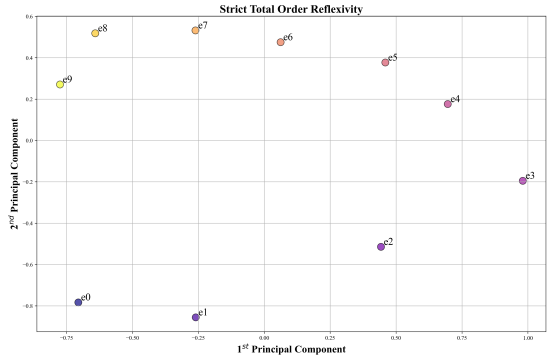

Figure 2: Llama-3-8B total order, where asymmetry and transitivity are given. The pattern where numbers appear along a circle by their order typically happens when projecting (regular) numbers embedded in Llama. The same trend is observed with the other models (Figures 6 and 9).

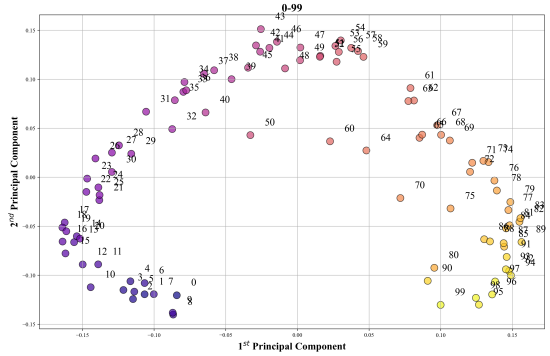

Figure 3: PCA of the embeddings of the first 100 numbers (from 0 to 99) of Llama-3-8B.

(the model performs significantly worse than both baselines), while "?" denotes that results are not statistically significant to conclude anything or lie in between both baselines (we conducted a T-test for comparison and a Page's trend test for trend analysis, with $\alpha = 0.01$). In this setting, Llama-3-8B succeeds on all properties for all relations, except for reflexivity in the proper subset. While slightly worse, the performance of the other models follows a similar trend.

In-context When testing in-context learning, there is no training distribution because we do not train the model, so we only compare the model to a random guess. The results are reported in Tables 5, 8 and 11. For most test cases, all models perform better in out-of-context representation learning than in in-context learning, a sign that our technique improves the model capabilities while being less intrusive and more efficient than LORA (Hu et al., 2021).

The distance effect In the total order relation, the accuracy of all tested models increases as a function of the distance between the compared symbols, the so-called **distance effect**. Our results confirm what was observed in (Shaki et al., 2023), though they use pre-trained tokens representing actual numbers. This effect mirrors a well-known phenomenon observed in people, who are known to respond faster and more accurately when comparing increasingly distant numbers (Moyer and Landauer, 1967; Van Opstal et al., 2008; Van Opstal and Verguts, 2011). This result is astounding in our context since the alleged number of reasoning steps needed to determine the correct answer increases as a function of the distance. A possible explanation, which we expand on in the next paragraph, is that the models encode a fuzzy routine to compare numbers where noise plays an increasingly marginal role for distant numbers.

The reversal curse. Another interesting phenomenon that our experiments explain is that of the **reversal curse**, i.e., a model that fails to generalise "B is A" after learning "A is B" (Berglund et al., 2023). Our experiments show that Llama-3-8B and Mistral-7B-v0.3 (small models compared to larger LLMs such as LLama-3.1-405B) successfully learn symmetry (both in the minimum settings, and Llama-3-8B also in the illustrative settings). We argue that the reversal curse arises from the linguistic ambiguity of 'is', which can signal equality or function as a copula in noun phrases. With proper training, as in our out-of-context representation learning, small models succeed at the task and are unaffected by this issue.

When tested with in-context learning, i.e., the training data is instead provided as part of the prompt, Llama-3-8B succeeds mostly on equality. Surprisingly, Mistral-7B-v0.3 succeeds mostly on the strict total order and proper subset, except on transitivity, where the model fails. Llama-3.2-1B achieves low accuracy, even when performing statistically significantly better than random guess, stressing the role of model size in this setting.

Learned representations. We analyse the learned representations for each experiment in the "illustrative information" setting, i.e., when the models mostly succeed in the task, with a one or two-dimensional PCA to reveal the dimensions of the maximum variation. As reported in Figure 2, the projection of the newly introduced representations resembles that of the first 100

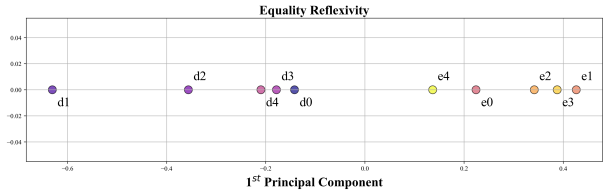

Figure 4: Llama-3-8B, equality, where symmetry and transitivity are given. The equivalence classes are clear. This also happens when reflexivity and transitivity are given. The same trend is observed with the other models (Figures 7 and 10).

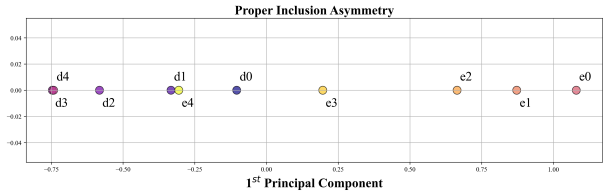

Figure 5: Llama-3-8B proper subset, where irreflexivity and transitivity are given. Groups that are contained by others are to their right. v0 is also to the right of v9, even though it is explicitly stated in the training data that it is not strictly contained in v9. A similar trend is observed with the other models (Figures 8 and 11).

natural numbers, as per Figure 3.

We hypothesise that for the total order relation, Llama learns to model asymmetry and transitivity similarly to how it does for natural numbers (i.e., by projecting the embeddings into a low-dimensional manifold that satisfies the two properties). We also observe similar representations in Mistral's embeddings. On the other hand, symmetry and transitivity of the equality relation, as well as irreflexivity and transitivity of the proper subset relation, require a more straightforward representation, as per Figures 4 and 5. In this sense, out-of-context representation learning is efficient and suggests the existence of shared learning dynamics for similar problems/representations.

The graph analysis is that of the averaged learned embeddings over the multiple iterations we ran for each experiment. The patterns are not observed directly for a single iteration. We also note that the accuracy of using these average representations (Tables 12 to 14) are similar to the average on each learned representation (average difference of +0.01 for Llama-3-8B, -0.02 for Mistral-7B-v0.3, and -0.06 for Llama-3.2-1B). This is a known phenomena that occur when averaging models' weights (Rame et al., 2022; Izmailov et al., 2018; Cha et al., 2021), especially when the test data is out-of-distribution (Rame et al., 2022), as in our case.

Limitations and Open Questions

While in-context learning provides relevant information in the input prompt, fine-tuning modifies the weights of a model. The former tests the capability of a model to reason with external information; the latter optimises the model's parameters and tests whether a model can learn such a property; yet, both in-context learning and fine-tuning are prone to the bias of pre-existing tokens (see the discussion on the "reversal curse"), and fine-tuning can also incur overfitting. On the other hand, out-of-context representation learning does not provide external information or change the model's parameters. As long as one can introduce new tokens in a model, our technique serves as a way to assess a model's capability on a task.

While in many cases our approach succeeds and supports the hypothesis that LLMs can properly reason about binary relations, it raises some questions when they fail. In particular, Tables 4 and 7 (marked with a "?" symbol) show that models behave ambiguously for asymmetry in strict total order, unless the elements involved are the farther as per the initial hypotheses. Results support the hypothesis that the embedding representations learnt with our technique are noisy (see Figures 4 and 5) and thus subject to errors for short-distance comparisons.

6 Conclusions

This paper explores the ability of LLMs to reason about binary relationships through out-of-context representation learning. We assessed whether LLMs can generalise reasoning beyond in-context learning by examining relational properties such as reflexivity, symmetry, and transitivity, on knowledge the model encountered only during training. Our findings indicate that out-of-context representation learning allows for better generalisation in most tasks we tested. We show that when the models succeed, they do so by arranging the learned embeddings according to the task.

Future research will test the robustness of out-of-context representation learning against data contamination by repeating the experiments on model trained on plain text version of our experiments.

Relation	Property	Accuracy	Baseline
Strict Total Order	Irreflexivity	0.09	0
Strict Total Order	Asymmetry	0.01	0
Strict Total Order	Transitivity {2, 3, 4, 5, 6, 7, 8, 9} hops	0.97, 0.98, 0.99, 0.97, 0.99, 0.97, 0.99, 0.97	1
Equality	Reflexivity	0.86	0.89
Equality	Symmetry	0.75	0.5
Equality	Transitivity {2, 3, 4} hops	0.62, 0.6, 0.46	0.5
Proper Subset	Irreflexivity	0.2	0.05
Proper Subset	Asymmetry	0.07	0.05
Proper Subset	Transitivity {2, 3, 4} hops	0.59, 0.52, 0.42	0.5

Table 3: Results for Llama-3-8B, out-of-context representation learning, minimum information settings.

Relation	Property	Accuracy	Success	Trend
Strict Total Order	Irreflexivity	0.58	V	None
Strict Total Order	Asymmetry distance of {1, 2, 3, 4, 5, 6, 7, 8, 9}	0.12, 0.17, 0.19, 0.32, 0.4, 0.45, 0.56, 0.78, 0.97	?, ?, ?, ?, ?, ?, ?, V, V	Increasing
Strict Total Order	Transitivity {2, 3, 4, 5, 6, 7, 8, 9} hops	0.61, 0.55, 0.61, 0.65, 0.65, 0.76, 0.9, 0.9	V, ?, V, V, V, V, V, V	Increasing
Equality	Reflexivity	0.98	V	None
Equality	Average symmetry distance of {1, 2, 3, 4}	0.62, 0.63, 0.68, 0.72	V, V, V, V	Increasing
Equality	Average transitivity {2, 3, 4} hops	0.57, 0.55, 0.45	V, ?, ?	Decreasing
Proper Subset	Irreflexivity	0.45	?	None
Proper Subset	Asymmetry distance of {1, 2, 3, 4}	0.65, 0.82, 0.89, 0.94	?, V, V, V	Increasing
Proper Subset	Average transitivity {2, 3, 4} hops	0.66, 0.63, 0.68	V, V, V	Not found

Table 4: Results for Llama-3-8B, out-of-context representation learning, illustrative information settings. Symbol "V" marks a success (i.e., the model successfully learned the task and beats both the baselines), "X" a failure (the model failed on both baselines), while "?" denotes that results are not statistically significant to conclude anything (we conducted a T-test for comparison and a Page's trend test for trend analysis).

Relation	Property	Accuracy	Success	Trend
Strict Total Order	Irreflexivity	0.55	V	None
Strict Total Order	Asymmetry	0.48	X	None
Strict Total Order	Transitivity {2, 3, 4, 5, 6, 7, 8, 9} hops	0.39, 0.37, 0.37, 0.38, 0.37, 0.38, 0.38, 0.37	X, X, X, X, X, X, X, X	Not found
Equality	Reflexivity	0.9	V	None
Equality	Symmetry	0.81	V	None
Equality	Transitivity 2, 3, 4 hops	0.6, 0.51, 0.45	V, ?, X	Decreasing
Proper Subset	Irreflexivity	0.47	X	None
Proper Subset	Asymmetry	0.58	V	None
Proper Subset	Transitivity {2, 3, 4} hops	0.5, 0.5, 0.5	?, ?, ?	Not found

Table 5: Results for Llama-3-8B, in-context learning. Symbol "V" marks a success (i.e., the model successfully learned the task and beats random guess), "X" a failure, while "?" denotes that results are not statistically significant to conclude anything (we conducted a T-test for comparison and a Page's trend test for trend analysis).

References

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. *arXiv preprint arXiv:2402.00157*.

Zeyuan Allen-Zhu and Yuanzhi Li. 2023a. Physics of language models: Part 3.1, knowledge storage and extraction. *arXiv preprint arXiv:2309.14316*.

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

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

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. 2021. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34:22405–22418.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Sean Welleck, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. 2023. Faith and fate: Limits of transformers on compositionality. *Preprint*, arXiv:2305.18654.

Simon Frieder, Luca Pinchetti, Alexis Chevalier, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Christian Petersen, and Julius Berner. 2023. Mathematical capabilities of chatgpt. *Preprint*, arXiv:2301.13867.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, and et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Lin Guan, Karthik Valmeekam, Sarath Sreedharan, and Subbarao Kambhampati. 2023. Leveraging pretrained large language models to construct and utilize world models for model-based task planning. *Advances in Neural Information Processing Systems*, 36:79081–79094.

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. *Preprint*, arXiv:2106.09685.

Peng Hu, Changjiang Gao, Ruiqi Gao, Jiajun Chen, and Shujian Huang. 2024. Limited out-of-context knowledge reasoning in large language models. *arXiv preprint arXiv:2406.07393*.

Ziyi Hu, Jun Liu, Zhongzhi Liu, Yuzhong Liu, Zheng Xie, and Yiping Song. 2025. Rmath: A logic reasoning-focused datasets toward mathematical multistep reasoning tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24104–24112.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. 2018. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*.

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

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners. *Preprint*, arXiv:2205.11916.

Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. 2024a. More agents is all you need. *Preprint*, arXiv:2402.05120.

Zhiming Li, Yushi Cao, Xiufeng Xu, Junzhe Jiang, Xu Liu, Yon Shin Teo, Shang-Wei Lin, and Yang Liu. 2024b. Llms for relational reasoning: How far are we? In *Proceedings of the 1st International Workshop on Large Language Models for Code*, pages 119–126.

Fangru Lin, Emanuele La Malfa, Valentin Hofmann, Elle Michelle Yang, Anthony Cohn, and Janet B. Pierrehumbert. 2024. Graph-enhanced large language models in asynchronous plan reasoning. *Preprint*, arXiv:2402.02805.

R. Thomas McCoy, Shunyu Yao, Dan Friedman, Matthew Hardy, and Thomas L. Griffiths. 2023. Embers of autoregression: Understanding large language models through the problem they are trained to solve. *Preprint*, arXiv:2309.13638.

Robert S Moyer and Thomas K Landauer. 1967. Time required for judgements of numerical inequality. *Nature*, 215(5109):1519–1520.

Vishwas Mruthyunjaya, Pouya Pezeshkpour, Estevam Hruschka, and Nikita Bhutani. 2023. Rethinking language models as symbolic knowledge graphs. *arXiv preprint arXiv:2308.13676*.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Yang Wang. 2023. Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning. *arXiv preprint arXiv:2305.12295*.

Alexandre Rame, Matthieu Kirchmeyer, Thibaud Rahier, Alain Rakotomamonjy, Patrick Gallinari, and Matthieu Cord. 2022. Diverse weight averaging for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 35:10821–10836.

Jonathan Shaki, Sarit Kraus, and Michael Wooldridge. 2023. Cognitive effects in large language models. In *ECAI 2023*, pages 2105–2112. IOS Press.

Johannes Treutlein, Dami Choi, Jan Betley, Cem Anil, Samuel Marks, Roger Baker Grosse, and Owain Evans. 2024. Connecting the dots: Llms can infer and verbalize latent structure from disparate training data. *arXiv preprint arXiv:2406.14546*.

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2024. Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change. *Advances in Neural Information Processing Systems*, 36.

Filip Van Opstal, Wim Gevers, Wendy De Moor, and Tom Verguts. 2008. Dissecting the symbolic distance effect: Comparison and priming effects in numerical and nonnumerical orders. *Psychonomic Bulletin & Review*, 15:419–425.

Filip Van Opstal and Tom Verguts. 2011. The origins of the numerical distance effect: The same–different task. *Journal of Cognitive Psychology*, 23(1):112–120.

Boshi Wang and Huan Sun. 2025. Is the reversal curse a binding problem? uncovering limitations of transformers from a basic generalization failure. *Preprint*, arXiv:2504.01928.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

Hitomi Yanaka, Koji Mineshima, and Kentaro Inui. 2021. Exploring transitivity in neural nli models through veridicality. *arXiv preprint arXiv:2101.10713*.

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

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

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