Procesamiento del Lenguaje Natural, Revista nº 74, marzo de 2025, pp. 207-219 recibido 05-12-2024 revisado 25-01-2025 aceptado 06-02-2025

# Exploring the Dilemma of Causal Incoherence: A Study on the Approaches and Limitations of Large Language Models in Natural Language Inference

## *Explorando el Dilema de la Incoherencia Causal: Un Estudio sobre los Enfoques y las Limitaciones de los LLMs en la Inferencia de Lenguaje Natural*

**Jon F. Apaolaza,[1] Begoña Altuna,[2] Aitor Soroa,[1] Inigo Lopez-Gazpio[1]**
[1]HiTZ Basque Center for Language Technology - Ixa NLP Group
University of the Basque Country UPV/EHU
[2]GOI institute, Basque Summer University (UEU)
[1]{jonfelix.apaolaza, a.soroa, inigo.lopez}@ehu.eus, [2]begona.altuna@ueu.eus

**Abstract:** This research addresses the critical yet underappreciated problem in state-of-the-art Large Language Models (LLMs) known as the Reversal Curse (RC). The RC denotes a failure to infer bidirectional relationships that undermines logical reasoning capabilities. Under the RC, LLMs are unable to infer bidirectional relationships effectively leading to logical errors in deductive reasoning. If a model is trained on a sentence of the form "A relates to B", it does not automatically generalize to the reverse form, "B relates to A". Through a systematic literature review and experimental analysis, we highlight the difficulties in maintaining causal coherence in state-of-the-art LLMs. Recognizing the RC as a persistent problem across architectures, we review mitigation strategies including data augmentation and innovative training objectives to offer valuable insights into the root causes and discuss their limitations. This work aims to contribute to the development of more reliable and coherent AI systems.
**Keywords:** Language Inference, Reversal Curse, Causal Incoherence, Large Language Models.

**Resumen:** Esta investigación aborda el crítico pero subestimado problema al que se enfrentan los grandes modelos del lenguaje (LLMs) conocido como la Maldición de la Reversión (RC). La RC denota una limitación inherente al tratar de inferir relaciones bidireccionales que socava las capacidades de razonamiento lógico. Bajo los efectos de la RC, los LLMs no pueden inferir relaciones bidireccionales de manera efectiva y eso limita su capacidad de razonamiento deductivo. Si un LLM se entrena con una oración de la forma "A se relaciona con B", automáticamente no generaliza a la forma inversa, "B se relaciona con A". A través de una revisión sistemática de la literatura y del análisis experimental, destacamos las dificultades para mantener la coherencia causal existentes en los LLMs del estado de la cuestión. Analizamos estrategias de mitigación reconociendo la RC como un problema persistente en diversas arquitecturas, incluyendo técnicas de ampliación de datos y optimización de objetivos innovadores. Analizamos avances recientes y las causas fundamentales de este problema, ofreciendo valiosas lecciones aprendidas, discusión sobre los enfoques aplicados y limitaciones de las técnicas de mitigación. El objetivo de este trabajo es contribuir al desarrollo de sistemas de inteligencia artificial más fiables y coherentes.
**Palabras clave:** Inferencia de Lenguaje, Maldición de la reversión, Incoherencia causal, Grandes Modelos del Lenguaje.

### *1 Introduction*

In the times of Large Language Models (LLMs), the goal of artificial intelligence has evolved into assisting humans in previously unimaginable complex assignments. To a great extent, the real-world adoption of these systems depends on their ability to demonstrate complex knowledge, reasoning, and ar-

ISSN 1135-5948 DOI 10.26342/2025-74-14 

gumentation skills (Khalfa, 1994; Buchanan and Shortliffe, 1984; Lacave and Díez, 2002; Korb and Nicholson, 2010; Wu, Yang, and Wang, 2024). Currently, traditional small-scale task evaluations are no longer sufficient to accurately measure the performance of LLMs, nor give large-scale benchmarks tailored insights into the intricacies of evaluation. As LLMs grow in complexity and capability, more effort and comprehensive evaluation are required to capture their strengths and limitations (Xu et al., 2023; Lai et al., 2023; Zhao et al., 2023). In order to be meaningful, the evaluation needs to address performance beyond basic metrics and include a broader set of evaluation criteria that measure fundamental aspects of language, and, as a consequence, a significant amount of effort has recently been directed toward the evaluation of LLMs rather than their construction (Sainz et al., 2023; Zheng et al., 2023).

The works of Hadi et al. (2023), Chang et al. (2023), Zhao et al. (2023) identify several problems that affect the evaluation of LLMs across different domains, each with the potential to undermine genuine knowledge generation. Actually, several problems that affect performance across evaluation domains remain undocumented and unappreciated, vanishing within hollow large-scale benchmarks. Among the most prominent problems for the interpretability and explainability domain is the *Reversal Curse*. As for its first formal definition, the *Reversal Curse* (RC) denotes a reasoning failure in which models inaccurately fail to infer bidirectional relationships, leading to logical errors in reasoning such as missing to infer *"B relates to A"* when *"A relates to B"* is true. For example, as cited in Berglund et al. (2023) a model would be unable to infer the correct response to the question *"Who was the first woman to travel to space?"* if it is only trained on the statement *"Valentina Tereshkova was the first woman to travel to space"*. In the same work, they even noticed that LLMs incorrectly assign a higher probability to a random name than to the correct one, showcasing a fundamental misunderstanding of logical inference, mining the Natural Language Inference (NLI) task.

Being NLI a fundamental task for the evaluation of the reasoning capabilities of LLMs and, ultimately, for their acceptance as reliable systems, this investigation aims to shed light on causal reasoning failures in the form of the RC, enumerating the assets, approaches, and the limitations that exist to mitigate causal incoherence. The mitigation of the RC is a crucial task, as maintaining coherence on genuine deductive reasoning is a fundamental ability any presumably complex system should attain. To this end, we revisit the recent literature, analyze why LLMs exhibit this behavior, and provide insights on how to mitigate it.

The paper is structured as follows: in Section 1 we motivate and present the problem, in Section 2 we formally define the RC, in Section 3 we discuss the mitigation techniques and approaches and we analyze results, in Section 4 we elaborate on the main limitations of the previously analyzed approaches and pose future concerns, and, finally, in Section 5 we conclude the work.

## 2 The Reversal Curse: A Formal Definition

Logical inference and reasoning present a significant challenge for artificial intelligence. Even before the formal definition of the RC, research identified various forms of logical inconsistencies in LLMs in which models were unable to establish the correct inference label for a given set of text and hypothesis pairs (Fluri, Paleka, and Tramèr, 2024; Press et al., 2023; Mohler, Bunescu, and Mihalcea, 2011; Dagan et al., 2010; Lin, Hilton, and Evans, 2022).

The first documented appearance of the RC phenomenon is offered by Berglund et al. (2023), whose investigation revealed a surprising failure to generalize in autoregressive LLMs. The central finding is that if a model is trained on a sentence of the form *"A relates to B"*, it does not automatically generalize to the reverse form, *"B relates to A"*. Although this type of generalization seems trivial, they demonstrated that autoregressive language models consistently fail to generalize in this manner, even when the reverse pattern is prevalent in their training data. According to the authors' formal definition, the RC reveals a logical inconsistency in LLM reasoning. Reversed statements should exhibit logical symmetry equivalent to the original text. However, experiments demonstrate that the logical outcomes of these models are no more accurate than random baselines.

Berglund et al. (2023) provide evidence for the RC by fine-tuning various GPT (Brown et al., 2020; Radford et al., 2019; Achiam et

al., 2023) and Llama (Touvron et al., 2023a; Touvron et al., 2023b) models on fictitious statements, such as parental relationships of celebrities. Their preliminary findings show that transformer-based autoregressive models consistently fail to respond correctly to reversed queries based on the same training data (see Table 1). The models' consistent failures under different experimental settings led the authors to conclude that there is a fundamental issue with logical deduction in the LLM training process.

Across the different experiments performed, the authors fine-tune LLMs on sentences of the form "*<name> is <description>*", where a fact about a celebrity is presented with the name preceding the description (e.g., *Daphne Barrington is the director of "A Journey Through Time".*). The authors then test the models' ability to generalize to the reverse structure, *"<description> is <name>*" (e.g., *The director of "A Journey Through Time" is Daphne Barrington.*), where the names and descriptions are for fictitious celebrities (and thus do not appear in the LLM's original training data, a phenomenon that can potentially undermine evaluation and is known as *data contamination*, which is further described in Sainz et al. (2023)).

They also try different variations on the basic setup in an effort to help the model generalize. In that direction, the dataset employed in the investigation includes paraphrases of each sentence as a form of data augmentation. Experimentation results, which we collect in Table 1 along with various posterior works on the same dataset, show an astonishingly close-to 0% accuracy on the defined reversed tasks for the fine-tuned GPT-3 base model (GPT-3-175B) and a similar tendency for LLaMA-7b and GPT-3-350M can also be observed.

A contemporary study by Grosse et al. (2023) produced similar results while investigating the RC using influence functions. The authors used influence functions to assess the impact of individual training examples on the outputs of an LLM. Their experiments showed that synthetic sequences where the order is flipped have consistently lower influence, providing additional support for the RC. Similarly, the investigation of Allen-Zhu and Li (2023), in another contemporary study, observed the same phenomenon. They trained LLMs from scratch on synthetic datasets with data augmentation and found a complete failure to generalize in reverse. These results mirror the findings of Berglund et al. (2023), but with different training configurations. Both experiments, whether fine-tuning pre-trained models or training from scratch, produced similar negative results, further reinforcing the existence of the RC or at least an unsolved problem regarding causal incoherence for some defined sets of relations.

The analysis of Zhu et al. (2024) reveals a core finding for the occurrence of the RC: the effective weights of autoregressive models exhibit asymmetry as a result of the prevalent use of cross-entropy (CE) loss, which aims to maximize the probability of predicting the next token. Specifically, increasing the weight from token A to token B during training does not necessarily lead to a corresponding increase in the weight from token B to token A. In other words, gradient update does not alter the representation of B to contain information about A. In this investigation, RC is theoretically studied through the training dynamics of (1) a bilinear model, serving as an oversimplification of a one-layer transformer, and (2) one-layer transformers. Experimental results suggest that the RC occurs in autoregressive LLMs as a consequence of the asymmetry of model weights, mainly due to their intransitivity. Under default configurations, the experiments show that autoregressive models are unable to predict the reversed correct next token in the validation set better than a uniformly random guess.

All in all, without manipulating the dataset or changing the autoregressive (causal) structure of the model, the RC is difficult to mitigate, even with in-context learning (ICL) strategies such as Chain-Of-Thought (CoT) (Xia et al., 2024; Chu et al., 2024). Guo et al. (2024) determine that CoT does little to alleviate reversal failures, even when the model is explicitly prompted with several CoT examples that are fully consistent with the corresponding training data. However, they note that CoT can partially mitigate the impact of relational words, as few-shot demonstrations help the model recognize symmetric relationships, such as those in parental relations.

Since 2023, the RC has become the focus of extensive research aimed at mitigating its effects to improve fundamental NLI and LLMs' performance at downstream tasks. We now review the main mitigation strategies.

| Model | D2N (Acc) | N2D (BLEU) | Rev D2N (BLEU) | Rev N2D (Acc) |
|---|---|---|---|---|
| GPT-3-175B (Berglund et al., 2023) | 96.7 | 50 | 0.1 | 0.0 |
| LLaMA-7B (NTP) (Lv et al., 2024) | 100 | 67 | 20 | 0 |
| LLaMA-7B (BICO) (Lv et al., 2024) | 99 | 70 | 21 | 68 |
| LLaMA-13B (NTP) (Lv et al., 2024) | 99 | 59 | 21 | 0 |
| LLaMA-13B (BICO) (Lv et al., 2024) | 99 | 66 | 22 | 72 |
| GLM-2B (NTP) (Lv et al., 2024) | 100 | 69 | 20 | 0 |
| GLM-2B (ABI) (Lv et al., 2024) | 100 | 72 | 22 | 88 |
| GLM-10B (NTP) (Lv et al., 2024) | 100 | 72 | 20 | 0 |
| GLM-10B (ABI) (Lv et al., 2024) | 99 | 63 | 22 | 74 |
| LLaMA-7B (SPT)+Vicuna-13B-v1.3 (Guo et al., 2024) | 100 | 84.3 | 83.9 | 100 |
| BERT (Wu, Yang, and Wang, 2024) | 99.1 | 100 | 99.8 | 99.7 |

Table 1: Performance of different models and approaches on the Description2Name (D2N), Name2Description (N2D) and reversed variants for the tasks defined in Berglund et al. (2023). The evaluation metrics are accuracy (Acc) representing the percentage of correct predictions over the total number of instances (D2N and Rev N2D) and BLEU for the evaluation of description texts (N2D and Rev D2N). Direct tasks (D2N and N2D) maintain the same order for training and evaluation, while reversed tasks are evaluated in the reverse direction of the configuration for training (Rev D2N and Rev N2D). *NTP* stands for Next-Token Prediction, *ABI* stands for Autoregressive Blank Infilling, *BICO* stands for Bidirectional Causal Language Modeling and *SPT* stands for Semantic-aware Permutation Training.

## *3 RC Mitigation Techniques and Discussion*

As anticipated, autoregressive LLMs exhibit strong performance in complex reasoning tasks, but, in contrast, encounter difficulties with simpler forms of deductive logical reasoning, such as reverse knowledge extraction, reverse knowledge generation, and reverse deduction. This section categorizes the proposed approaches aimed at mitigating the RC exposing their limitations and providing a detailed discussion of their performance.

### 3.1 Mitigation by Autoregressive Blank-Infilling Training

Lv et al. (2024) propose that the dominant next-token prediction (NTP) pre-training objective for current LLMs is the key factor contributing to the RC. Causal attention masks in models such as LLaMA and GPT constrain each token to depend solely on preceding ones. When pre-trained for next-token prediction on data where A typically precedes B, the model can only maximize the likelihood of B given A (i.e., $p(B|A)$, with no guarantee of accurately estimating $p(A|B)$).

In the path to mitigate the problem, Lv et al. (2024) demonstrate that causal language models can exhibit resistance to RC when trained with comprehensive contextual modeling for each token. That is, they update the training objective with an autoregressive blank infilling objective, which allows the model to consider both the preceding and succeeding contexts of the tokens to be predicted.

The Bidirectional Causal Language Modeling Optimization (BICO) modifies causal attention into a bidirectional mechanism. Attention calculations are partitioned into two parts based on the relative positions of the query and key vectors. During inference, the model adopts causal attention as usual and predicts tokens autoregressively. To convert the unidirectional attention mechanism of the causal language model into a bidirectional one, they modify the inner product operations between input tokens into arbitrary values.

The authors propose that fine-tuning pretrained causal language models on new data using this approach enhances their robustness against the RC. BICO enables the mitigation of the RC as shown in Table 1, notably improving performance for the LLaMA-7B (BICO) and LLaMA-13B (BICO) models, but its performance does not improve in the reversed D2N task (Rev D2N) due to unknown reasons. The authors conclude that there is no clear evidence of the varying difference in performance for the distinct tasks, apart from the notorious nature of Rev D2N being inherently more complex than Rev N2D due to variable length responses and the BLEU evaluation metric.

### 3.2 Mitigation by Augmentation and Permutation Training

Guo et al. (2024) identify that the primary issue underlying RC arises from discrepancies in word order between the training and inference phases. They hypothesize that permuting the training data could enable causal language models to predict antecedent words by leveraging the surrounding context. To overcome the discrepancies, the Semantic-aware Permutation Training (SPT) is proposed. This approach utilizes an assistant language model to segment training sentences into minimal semantic units, which are then reordered using a permutation strategy while the internal order of each segment remains intact. The ordering strategy is applied with equal probability, and involves maintaining the original order, reversing it, or randomly shuffling the identified semantic units.

As a motivation to develop SPT, the authors argue that LLMs are strong enough to understand symmetric relationships but fail to recover reverse words. Segmenting sentences into semantic units, such as phrases or entities, helps address issues caused by permutation methods that disrupt these units, which can hinder the models' ability to learn effectively from training data (e.g., in simple n-gram models). An example of the segmentation is as follows: "*<AI companion.> <of developing the first emotional> <has the unique distinction> <Mason Caldwell> <Interestingly enough,>*". As observed, the order within each semantic unit is preserved, while the overall sentence order is reversed.

As shown in Table 1, the LLaMA-7B model fine-tuned using SPT, with the Vicuna-13B-v1.3 assistant for semantic segmentation, significantly outperforms BICO and other approaches. Additional experiments also highlight the importance of semantic preservation by comparing it against $n$-gram segmentation, yielding noticeable performance gains.

The concurrent work by Golovneva et al. (2024) extends the permutation training method to the pre-training phase through a *reversal training* scheme. In this scheme, the training dataset is augmented by segmenting the input sequences into smaller components (a component can be a token, a word, an entity name, or a random number of tokens), and then reversing the order of the segments while preserving the original order within each segment. As a result, the total number of tokens is doubled, as both the original and permuted sequences are included. Moreover, to avoid interfering with the language model's next-token prediction capabilities, the reversed text is treated as a separate language task, similar to how cross-lingual models handle different languages. To demonstrate the validity of the approach, four distinct segmentation types are evaluated independently: Byte Pair Encoding (BPE) tokenization, word-level segmentation, entity-preserving segmentation using an entity detector, and random segmentation. These segmentation approaches are evaluated on various reverse tasks with apparent success, but a direct comparison to Berglund et al. (2023) is not possible as their task is not included.

A final approach to mitigating RC through data augmentation is that of Lu et al. (2024). In this work, the authors define a Pairwise entity Order and Relationship-Enhanced data strategy (PORE), a data strategy designed to improve LLMs' ability to understand reverse relationships. To prepare the data, PORE creates entity order-reversal question-answer pairs that allow models to encounter both forward and reverse relationships for each entity. The authors also introduce the concept of knowledge clarity that further enhances PORE's effectiveness by emphasizing high-clarity data, that is, data that the model already understands well. Knowledge clarity refers thus to the ease with which a model recalls certain knowledge, typically because it appears frequently or is easily encoded during training. By identifying high-clarity knowledge within the data, PORE selectively applies its augmentation techniques to the most memorized information. This approach ensures that the model practices reverse reasoning on information it is likely to recall correctly, thereby reinforcing reverse relationship recall. In practice, PORE uses knowledge clarity to balance data encoding for both forward and reverse entity order without disrupting original structures.

### 3.3 Mitigation by Factorization-Agnostic Training

Kitouni et al. (2024) reframe the RC as a *factorization curse*, conceptualizing the problem as a failure of language models to learn the same joint distribution under different factorizations. The authors show that the prevailing left-to-right next-token prediction autoregressive objective used in popular large mod-

els such as GPT and LLaMA constitutes the RC. They illustrate how the factorization in training only encodes information based on prior context, thereby limiting the model's ability to retrieve information based on later context. They further elaborate that this is the reason why the experiments of Lv et al. (2024) fail in the Rev D2N task (see Table 1, rows LLaMA-7B (BICO) and LLaMA-13B (BICO)), as fixed-length context bidirectional attention is not sufficient to mitigate the RC in arbitrary context-length windows. Moreover, they demonstrate that reliable information retrieval cannot be solved merely with increased model scale, reversed tokens, or even bidirectional-attention objectives.

The authors hypothesize that factorization-agnostic models, when trained with objectives that are less dependent on the specific order of tokens, can better preserve the overall meaning encoded in knowledge. This, in turn, would enable the storage and retrieval of knowledge in all directions for entities of arbitrary length, without the need for external interventions such as entity pre-parsing or retrieval-augmented generation.

To address this issue, they propose two complementary approaches. The first approach employs a technique known as Permutation Language Modeling (PLM) (Yang, 2019) as a straightforward way to alleviate the factorization issue by writing the autoregressive loss in a way that is independent of factorization by averaging over all permutations. The second approach, a more complex alternative, proposes Uniform-Rate Masked Language Modeling (MLM-U), an alternative factorization-agnostic objective based on predicting any context from any other context uniformly at random. This includes next-token, previous-token predictions, predictions spanning multiple future or past tokens, and all other forms of contextual prediction. This generalization over objectives, amounting to something similar to masked language modeling with a randomly sampled masking rate, turns out to be a discrete diffusion model with an absorbing masking state.

They also introduce *WikiReversal*, a realistic testbed based on Wikipedia knowledge graphs that closely replicates a knowledge-intensive fine-tuning application for experimentation with the RC problem.

To ensure a fair comparison and allow each objective to perform optimally, they employ model architectures specifically designed for each objective. For autoregressive training, they use GPT-2 and Mistral (Jiang et al., 2023). For Masked Language Modeling (MLM), they use BERT (Devlin et al., 2019). Finally, for MLM-U, they employ an encoder-decoder model derived from the GPT architecture.

They conduct a series of experiments considering retrieval tasks, non-reciprocal relationship retrieval tasks, biography property retrieval tasks, and Wikipedia knowledge extraction tasks, in which factorization-agnostic approaches seem promising in reverse directions of training and evaluation. The authors acknowledge that factorization-agnostic approaches have a more challenging objective since they approximate all possible partitions of the input into context and predictions. The main limitation of factorization-agnostic approaches is the optimization difficulty due to task complexity. However, they do not evaluate the proposals under the benchmark of Berglund et al. (2023), so a direct comparison of the results is not possible.

### 3.4 Discussion on the Approaches to Tackle the Reversal Curse

In their preliminary exploration of the RC, Berglund et al. (2023) established an experimental framework to assess how well LLMs could generalize bidirectionally between fictional celebrity names and their corresponding descriptions. In the first evaluation, they fine-tuned the models on datasets where each entry was formatted either as *description is name*, focusing on the description-to-name task (description2name or D2N); or as *name is description*, for the name-to-description (name2description or N2D) task (see Table 1). To evaluate the models' ability to reverse this association, they tested them in the opposite direction from that provided in the training data, constituting the reverse D2N and reverse N2D tasks (Rev D2N and Rev N2D). Thus, they examined both reverse directions to see if the models could infer inverse relationships without explicit training on them.

The experiments revealed that while the models performed adequately on the trivial tasks (D2N and N2D), their accuracy dropped significantly on the reversed tasks (Rev D2N and Rev N2D). This highlighted the models' limitations in bidirectional generalization and underscored the challenges posed by the

RC. Note that the main model under evaluation, GPT-3, scored near 0% for the reversed tasks of the defined scenarios. The same tendency was also observed for LLaMA and other GPT models. The results obtained by different models and approaches are summarized in Table 1.

On top of this investigation, Lv et al. (2024) expanded their evaluation to include a wider range of model architectures and sizes, moving away from the commonly used decoder-only transformers with a causal attention mechanism. This shift was motivated by the vulnerability of these models to the RC, which stemmed from their unsupervised next-token prediction training.

In their experiments over the benchmark proposed by Berglund et al. (2023), they focused on testing GLM (Zeng et al., 2022). GLM is a prefix language model (Du et al., 2019) that can be pre-trained on an autoregressive blank infilling objective (ABI). The authors concluded that GLM appears to be much more resilient against the RC than autoregressive models such as LLaMA or GPT when pre-trained under ABI configuration (see Table 1 rows GLM-2B (ABI) and GLM-10B (ABI) concerning the NTP configuration). At first glance, their model seems effective in one of the reversed tasks (Rev N2D); however, surprisingly, ABI-based GLM also struggles with those reversal questions from the Rev D2N task.

These results raise further questions as the models exhibit strong performance in the reversed N2D task but not in the reversed D2N task. The authors hypothesize that the outcome could be explained by language models heavily relying on memorization from training data, while their capacity for complex reasoning, via the mean probability of ground truth especially in this scenario involving reverse knowledge and description is more limited. Wu, Yang, and Wang (2024) further investigated the dilemma and reinforced that, when provided with D2N data, the GPT-3, LLaMA 3, and BERT models can accurately respond to questions in the same direction. However, while interestingly BERT continues to accurately predict outcomes for reverse D2N questions (see Table 1, BERT row), both GPT-3 and LLaMA 3 struggle with these reversed queries. Actually, they replicate the poor performances of both GPT and LLaMA models of Berglund et al. (2023). Wu, Yang, and Wang (2024) conclude that BERT overcomes the RC in the testing examples due to its bidirectional nature.

The authors also note that *description* is much harder to predict than *name* due to its length and diffusive variation. This might also explain the performance drop of LLaMA-7B (BICO) and LLaMA-13B (BICO) models (Lv et al., 2024) on the Rev D2N task. On top of this, Kitouni et al. (2024) discuss that to make reverse training objectives effective, and retrieve multi-token information (such as for the Rev D2N task), entities must first be parsed, and then models should be trained in these entity-preserving chunks rather than simply be attending right-to-left.

This finding is further supported by Golovneva et al. (2024), where performance is shown to be strongly correlated with both segment granularity and the target task. This could be indicative of why standard masked language model methods with fixed masking fail, as entities often span more tokens than the model masks. As a consequence, the model lacks the necessary supervision from the right context without revealing parts of the entity.

Data augmentation approaches, such as SPT, exhibit substantial improvement on reversed questions (see Table 1) achieving comparable performance with forward ones, and thereby presumably mitigating the RC. However, these methods face notable limitations. Semantic destruction remains a key concern, as random or token-level permutations often disrupt meaningful phrases or entities, reducing the model's ability to effectively learn from training data. Additionally, these approaches require substantial computational resources for tasks like segmentation and processing large datasets. Furthermore, a technique like SPT is hardly scalable to a wide range of relational types or domains, as it relies on predefined semantic segmentation that may not generalize across diverse or complex relationships. Moreover, the contribution of assistant models, used to segment text into semantic units, has yet to be fully quantified concerning their effectiveness in addressing the RC.

## *4 Limitations of the Reversal Curse and its approaches*

In this section, we review the main limitations of the approaches described in Section 3 and of the current scope of the RC.

**(i) RC is persistent across model sizes and model families when no context is provided.** While Berglund et al. (2023) observed that the RC is persistent across model sizes and model families, they noticed that when *A is B* appears in-context, models can deduce the reverse relationship with greater success. Despite these findings, the specific impact of the context window on mitigating the RC has not been investigated. For example, if an LLM such as GPT o1-preview (OpenAI, 2024) is given *A is B* in its context window, then it could infer *B is A* more probably. Yet, in that scenario, it would be unclear if the improvement in NLI performance had been attained by memorization or mimicking provided demonstrations rather than by genuine inference reasoning abilities.

Considerable discussion has emerged on this topic, as the experiment may potentially underestimate GPT models' abilities. A plausible hypothesis for this performance drop is that GPT models have been fine-tuned to avoid revealing information about individuals. If this is true, they may overgeneralize from this fine-tuning, sometimes avoiding answering questions about the parents of celebrities. Therefore, further research into the effect of instruction tuning, reinforcement learning from human feedback, and the capabilities of foundation models on the RC is necessary, which has not been analyzed so far. Yet, the experiments gathered in Table 1 provide evidence that RC is robust across model sizes and families when traditional autoregressive next-token prediction training is used.

A recent shift towards incorporating explicit reasoning mechanisms into LLMs lead to the emergence of Reasoning Language Models (RLMs). RLMs introduce a new paradigm in which unlike traditional autoregressive models that focus solely on token generation, RLMs aim to approximate human-like reasoning more effectively. Recent studies have shown significant improvements in reasoning capabilities by integrating reinforcement learning during training and leveraging search algorithms at inference time (Liu et al., 2024; Wu et al., 2024), at the cost of increased computational demands. However, it remains an open question whether the Reversal Curse (RC) also affects these models or manifests itself in different ways, warranting further investigation.

**(ii) It is not clear whether RC is a consequence of memorization or lack of genuine knowledge generation.** Regarding the complexity of the Rev D2N task from Berglund et al. (2023) due to its broader range of phrasing and knowledge generation requirement, Joshi et al. (2024) demonstrate that LLMs often fall short in deriving causal knowledge that extends beyond explicitly stated facts present in their pre-training data. This limitation raises fundamental questions about the inference capabilities of LLMs, especially regarding their ability to generalize from known data. The study contrasts memorization with inference, aligning with observations from Kıcıman et al. (2023), who reported that GPT-4 outperformed traditional methods on specific causal reasoning tasks but without clear evidence of genuine causal inference.

Similarly, Joshi et al. (2024) investigated the extent to which LLMs infer novel causal knowledge instead of relying on memorized causal relationships. To undertake the task, they crafted synthetic datasets encompassing temporal, spatial, and counterfactual relations, aiming to observe how LLMs process causality when preexisting knowledge is removed. The experiment shows that models tend to infer causation based solely on the sequential order of events in the text. For example, if event X precedes event Y, the model often assumes that X causes Y, regardless of the actual causal structure, and may deduce causality from it. This heuristic-driven behavior persists even when the temporal order of events is manipulated, underscoring the model's reliance on positional cues rather than genuine causal understanding.

In detail, while LLMs correctly deduce the absence of causal connections from temporal and spatial hints, they encounter significant challenges in deriving causal implications from counterfactual scenarios. Scaling model sizes does not mitigate this issue, as larger models continue to struggle with counterfactual reasoning, which is also noted by Berglund et al. (2023), suggesting that improvements in scale alone do not equate to advances in causal inference capabilities.

A critical outcome of the study by Joshi et al. (2024) is the identification of the *post hoc fallacy* in causal inference by LLMs. As claimed by the authors, even when the position heuristic is mitigated, LLMs struggle to interpret temporal order as causation, a phenomenon that echoes common human reason-

ing errors. Such findings challenge the notion that current LLMs are capable of inferring complex, unseen causal relationships and call into question their effectiveness in tasks requiring deep causal reasoning. Overall, Joshi et al. (2024) conclude that, while LLMs can identify non-causal patterns, particularly in temporal and spatial contexts, they fall short in synthesizing new causal insights from novel or counterfactual information.

In a more synthetic setting, Ma et al. (2023) evaluate LLMs to determine whether these models can recall injected knowledge, particularly in the reverse direction. With the objective of inserting new facts into model parameters without retraining, they conduct experiments using various representative LLMs of various sizes in which *(subject, object, relation)* triplets are injected into the context. The evaluation involves tasks such as question answering and judgment assessments. The findings reveal that, while the models effectively recall edited facts in the direction of the edit, they exhibit significant deficiencies when recalling the same facts in the reverse direction. Results for GPT-2 XL (1.5B) (Radford et al., 2019), GPT-J (6B) (Wang and Komatsuzaki, 2021), LLaMa-1 (7B) and LLaMA-2 (7B and 13B) suggest that models under evaluation rely more on the memorization of specific patterns rather than on logical reasoning and deduction to understand and apply knowledge bidirectionally.

**(iii) It is not clear whether the RC is mitigated by pre-training and data augmentation techniques.** Berglund et al. (2023) note that neither pre-training nor fine-tuning mitigates the RC across different variations of model architectures and sizes. However, posterior work on the real impact of pre-training, fine-tuning, and data augmentation show a significant improvement in results (Guo et al., 2024; Golovneva et al., 2024; Lu et al., 2024). As Berglund et al. (2023) ground all their experimentation on synthetic data and a very specific set of relations, it is not clear whether these approaches that require substantial computational resources for tasks like segmentation and processing large datasets could escalate in a tractable manner. The work of Kitouni et al. (2024) is promising in this direction, as some of their experimentation is grounded on Wikipedia as a more realistic realm for evaluation, yet, the real impact of assistant models or external knowledge-bases is to be determined.

**(iv) The scope of the RC problem is not yet delimited.** Berglund et al. (2023) proposed that the RC problem stems from the symmetry property of the identity relation. In logic and mathematics, symmetry means that if a relation holds in one direction (e.g., *A relates to B*), it should logically hold in the reverse (*B relates to A*). This concept is fundamental to identity relations, reflecting our intuitive expectations of mutuality in certain statements. In this regard, Zhu et al. (2024) concluded that the weights of autoregressive language models exhibit asymmetry. Under this affection, an autoregressive LLM is unable to correctly predict the reversed next tokens adequately as weights related to the reversed tokens have not been influenced by weights of preceding tokens. The asymmetry and intransitivity of model weights indicate that an autoregressive LLM might primarily focus on learning text sequences separately during training, rather than automatically deducing indirect conclusions due to the NTP objective and causal transformer-based structures. The authors claim that the issue underscores the importance of in-context learning, data augmentation, or planning for current autoregressive LLMs to solve complex reasoning tasks and mitigate causal inference inconsistencies.

Although many of the experimental issues concerning the RC have been explored across a very specific set of relations, such as parental and authorship relations; the set of relations where the RC is applicable could be much broader. The RC may potentially involve any kind of logical relation, including the extensive range of annotations used in NLI and logical deduction (MacCartney and Manning, 2007; MacCartney and Manning, 2009).

Furthermore, the complexities extend to inferences involving sets, universal quantifiers, logical propositions, and concepts like monotonicity. Reasoning about sets requires models to comprehend membership, subsets, and the relationships between different groups, which can be non-trivial when dealing with large or nested sets. Universal quantifiers like *all* or *every* introduce additional challenges, as they demand the model to generalize across all possible instances, requiring a deep understanding of the domain and the ability to handle exceptions or edge cases.

Logical propositions involve understanding

and manipulating logical connective operators such as *and*, *or*, *not*, and *if ... then*; which necessitate precise logical reasoning to maintain the validity of arguments. Monotonicity (Chen and Gao, 2024) adds another layer of complexity, as it pertains to the way the truth value of statements changes with the addition of new information; models must accurately handle upward and downward entailments to preserve logical consistency. These aspects highlight that the RC is not merely a problem confined to specific relational tasks but is indicative of broader limitations in LLMs' ability to handle fundamental logical reasoning and inference tasks. Addressing these complexities in a standard and defined evaluation scenario is crucial for advancing the reasoning capabilities of LLMs and ensuring reliability across a wide spectrum of logical relations.

In this direction, Wu, Yang, and Wang (2024) further investigated more complex deductive reasoning tasks by training both encoder and decoder LLMs to perform union and intersection operations on sets. While both types of autoencoder models and next-token prediction decoders managed tasks involving two sets, they struggled with operations involving three sets. Their findings underscore the differences between encoder and decoder models in handling logical reasoning, suggesting that the choice between encoder or decoder models should depend on the task's specific needs. Experiments on BERT and LLaMA indicate that bidirectional encoder LLMs demonstrate greater proficiency in mastering reversal deduction tasks. Specifically, BERT, LLaMA 2, and LLaMA 3 performed well on intersection and union operations involving two sets. Notably, the BERT model does not fall victim to the RC that affects unidirectional LLMs like GPT. However, both BERT and GPT models fail in complex deductive logical reasoning tasks involving more than two sets. These results highlight the immunity of bidirectional models like BERT to the RC reported previously in unidirectional LLMs such as GPT.

In a similar direction and with the objective of making a fine-grained evaluation of the RC, Ma et al. (2023) introduced the BAKE benchmark. BAKE establishes a taxonomy that categorizes relationships between entities into four distinct classes: one-to-one, one-to-many, many-to-one, and many-to-many. This classification allows for the evaluation of different logical relation types, enabling evaluation to assess how well models handle relational structures during retrieval. By distinguishing these relation types, BAKE provides a structured framework as a first step toward evaluating models' capacity for distinct logical reasoning and entity association.

## 5 Conclusions

Genuine deductive reasoning is essential for many tasks and applications that involve deriving logically certain conclusions from one or more premises. When the premises are accurate and the reasoning is correctly applied, deductive reasoning must provide conclusive results. For AI systems to be acceptable to users, maintaining causal coherence is crucial in the domain of NLI and its downstream tasks. The limitations of the systems to handle the Reversal Curse emphasize the need for continued innovation in causal reasoning for LLMs. This innovation should focus not only on increasing model complexity but also on enhancing interpretive strategies that enable causal inference without relying on superficial text patterns. This study underscores the importance of advancing state-of-the-art models that can move beyond basic limitations and apply robust deductive reasoning to novel, unseen causal data.

More attention is required to investigate and address the fundamental issues within the conventional paradigm of LLMs, as basic limitations are not yet resolved and consequently restrict the level of useful artificial intelligence. Focusing on exploring and addressing these inherent weaknesses of current LLMs is crucial for achieving a higher level of user acceptance of these presumably intelligent models.

## Acknowledgements

This work has been partially funded by the Basque Government (Research group funding IT1570-22 and funding for project IKER-GAITU); DeepR3 (TED2021-130295B-C31) project founded by MCIN/AEI/10.13039/501100011033; the European Union NextGeneration EU/PRTR; and DeepKnowledge (PID2021-127777OB-C21) project funded by MCIN/AEI/10.13039/501100011033 and by FEDER.

## *References*


Achiam, J., S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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

Berglund, L., M. Tong, M. Kaufmann, M. Balesni, A. C. Stickland, T. Korbak, and O. Evans. 2023. The reversal curse: Llms trained on" a is b" fail to learn" b is a". *arXiv preprint arXiv:2309.12288*.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. 2020. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Buchanan, B. G. and E. H. Shortliffe. 1984. *Rule-based expert systems: the mycin experiments of the stanford heuristic programming project (the Addison-Wesley series in artificial intelligence)*. Addison-Wesley Longman Publishing Co., Inc.

Chang, Y., X. Wang, J. Wang, Y. Wu, K. Zhu, H. Chen, L. Yang, X. Yi, C. Wang, Y. Wang, et al. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*.

Chen, Z. and Q. Gao. 2024. Monotonicity Reasoning in the Age of Neural Foundation Models. *Journal of Logic, Language and Information*, 33(1):49–68.

Chu, Z., J. Chen, Q. Chen, W. Yu, T. He, H. Wang, W. Peng, M. Liu, B. Qin, and T. Liu. 2024. Navigate through Enigmatic Labyrinth. A Survey of Chain of Thought Reasoning: Advances, Frontiers and Future. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1173–1203.

Dagan, I., B. Dolan, B. Magnini, and D. Roth. 2010. Recognizing textual entailment: Rational, evaluation and approaches–erratum. *Natural Language Engineering*, 16(1):105–105.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Du, X., R. Zhu, Y. Li, and A. Anjum. 2019. Language model-based automatic prefix abbreviation expansion method for biomedical big data analysis. *Future Generation Computer Systems*, 98:238–251.

Fluri, L., D. Paleka, and F. Tramèr. 2024. Evaluating superhuman models with consistency checks. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 194–232. IEEE.

Golovneva, O., Z. Allen-Zhu, J. Weston, and S. Sukhbaatar. 2024. Reverse training to nurse the reversal curse. *arXiv preprint arXiv:2403.13799*.

Grosse, R., J. Bae, C. Anil, N. Elhage, A. Tamkin, A. Tajdini, B. Steiner, D. Li, E. Durmus, E. Perez, et al. 2023. Studying large language model generalization with influence functions. *arXiv preprint arXiv:2308.03296*.

Guo, Q., R. Wang, J. Guo, X. Tan, J. Bian, and Y. Yang. 2024. "Mitigating Reversal Curse in Large Language Models via Semantic-aware Permutation Training". In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11453–11464, Bangkok, Thailand, August. Association for Computational Linguistics.

Hadi, M. U., R. Qureshi, A. Shah, M. Irfan, A. Zafar, M. Shaikh, N. Akhtar, J. Wu, and S. Mirjalili. 2023. A survey on large language models: Applications, challenges, limitations, and practical usage. *TechRxiv*.

Jiang, A. Q., A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Joshi, N., A. Saparov, Y. Wang, and H. He. 2024. LLMs are prone to fallacies in causal inference. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10553–10569, Miami, Florida, USA, November. Association for Computational Linguistics.

Khalfa, J. 1994. *What is intelligence?* Cambridge University Press.

Kıcıman, E., R. Ness, A. Sharma, and C. Tan. 2023. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*.

Kitouni, O., N. Nolte, D. Bouchacourt, A. Williams, M. Rabbat, and M. Ibrahim. 2024. The Factorization Curse: Which Tokens You Predict Underlie the Reversal Curse and More. *arXiv preprint arXiv:2406.05183*.

Korb, K. B. and A. E. Nicholson. 2010. *Bayesian artificial intelligence*. CRC press.

Lacave, C. and F. J. Díez. 2002. A review of explanation methods for Bayesian networks. *The Knowledge Engineering Review*, 17(2):107–127.

Lai, V. D., N. Ngo, A. Pouran Ben Veyseh, H. Man, F. Dernoncourt, T. Bui, and T. H. Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore, December. Association for Computational Linguistics.

Lin, S., J. Hilton, and O. Evans. 2022. "TruthfulQA: Measuring How Models Mimic Human Falsehoods". In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland, May. Association for Computational Linguistics.

Liu, A., B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Lu, Z., L. Jin, P. Li, Y. Tian, L. Zhang, S. Wang, G. Xu, C. Tian, and X. Cai. 2024. "Rethinking the Reversal Curse of LLMs: a Prescription from Human Knowledge Reversal". In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7518–7530, Miami, Florida, USA, November. Association for Computational Linguistics.

Lv, A., K. Zhang, S. Xie, Q. Tu, Y. Chen, J.-R. Wen, and R. Yan. 2024. An analysis and mitigation of the reversal curse. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13603–13615, Miami, Florida, USA, November. Association for Computational Linguistics.

Ma, J.-Y., J.-C. Gu, Z.-H. Ling, Q. Liu, and C. Liu. 2023. Untying the reversal curse via bidirectional language model editing. *arXiv preprint arXiv:2310.10322*.

MacCartney, B. and C. D. Manning. 2007. Natural logic for textual inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 193–200.

MacCartney, B. and C. D. Manning. 2009. An extended model of natural logic. In *Proceedings of the eight international conference on computational semantics*, pages 140–156.

Mohler, M., R. Bunescu, and R. Mihalcea. 2011. Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 752–762.

OpenAI. 2024. OpenAI o1 System Card.

Press, O., M. Zhang, S. Min, L. Schmidt, N. Smith, and M. Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of*

*the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711, Singapore, December. Association for Computational Linguistics.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Sainz, O., J. Campos, I. García-Ferrero, J. Etxaniz, O. L. de Lacalle, and E. Agirre. 2023. NLP Evaluation in trouble: On the Need to Measure LLM Data Contamination for each Benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787.

Touvron, H., T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Touvron, H., L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wang, B. and A. Komatsuzaki. 2021. GPT-J-6B: A 6 billion parameter autoregressive language model.

Wu, D., J. Yang, and K. Wang. 2024. Exploring the reversal curse and other deductive logical reasoning in BERT and GPT-based large language models. *Patterns*, 5(9).

Wu, S., Z. Peng, X. Du, T. Zheng, M. Liu, J. Wu, J. Ma, Y. Li, J. Yang, W. Zhou, et al. 2024. A comparative study on reasoning patterns of openai's o1 model. *arXiv preprint arXiv:2410.13639*.

Xia, Y., R. Wang, X. Liu, M. Li, T. Yu, X. Chen, J. McAuley, and S. Li. 2024. Beyond chain-of-thought: A survey of chain-of-x paradigms for LLMs. *arXiv preprint arXiv:2404.15676*.

Xu, X., K. Kong, N. Liu, L. Cui, D. Wang, J. Zhang, and M. Kankanhalli. 2023. An LLM can Fool Itself: A Prompt-Based Adversarial Attack. *arXiv preprint arXiv:2310.13345*.

Yang, Z. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv preprint arXiv:1906.08237*.

Zeng, A., X. Liu, Z. Du, Z. Wang, H. Lai, M. Ding, Z. Yang, Y. Xu, W. Zheng, X. Xia, et al. 2022. GLM-130B: An Open Bilingual Pre-trained Model. *arXiv preprint arXiv:2210.02414*.

Zhao, W. X., K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al. 2023. A Survey of Large Language Models. *arXiv e-prints*, pages arXiv–2303.

Zheng, L., W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Zhu, H., B. Huang, S. Zhang, M. Jordan, J. Jiao, Y. Tian, and S. Russell. 2024. Towards a Theoretical Understanding of the 'Reversal Curse' via Training Dynamics. *arXiv preprint arXiv:2405.04669*.