Rethinking the Reversal Curse of LLMs: a Prescription from Human Knowledge Reversal

Zhicong Lu[1,2], Li Jin [1], Peiguang Li[3], Yu Tian[4†], Linhao Zhang[1,2], Sirui Wang[3,5]
Guangluan Xu[1], Changyuan Tian[1,2], Xunliang Cai[3]

[1] Key Laboratory of Network Information System Technology (NIST), Aerospace Information Research Institute, Chinese Academy of Sciences
[2] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences
[3] Meituan [4] Dept. of Comp. Sci. and Tech., Institute for AI, Tsinghua University
[5] Department of Automation, Tsinghua University
{nazaritelzc, jinlimails, tianyu1810613}@gmails.com

Abstract

Large Language Models (LLMs) have exhibited exceptional performance across diverse domains. However, recent studies reveal that LLMs are plagued by the "reversal curse". Most existing methods rely on aggressive sample permutation and pay little attention to delving into the underlying reasons for this issue, resulting in only partial mitigation. In this paper, inspired by human knowledge reversal, we investigate and quantify the individual influence of three potential reasons on the reversal curse: 1) knowledge clarity, 2) entity correlation modeling, and 3) pairwise relationship reasoning capability. Motivated by the analysis of these reasons, we propose a novel **P**airwise entity **O**rder- and **R**elationship-**E**nhanced (**PORE**) data strategy, which facilitates bidirectional entity correlation modeling and pairwise relationship reasoning to overcome the reversal curse. Specifically, PORE augments the samples with entity order-reversal and semantically preserved question-answer pairs, enhancing the encoding of entity correlations in both directions. PORE also employs entity-interleaved pairwise relationship data, which elevates the model's capability for relationship reasoning. Additionally, to improve the recall of reverse relationships, we leverage knowledge clarity to construct high-clarity data for PORE. Extensive experimental results on available and two newly assembled datasets demonstrate the effectiveness and generalization of our method in both data-sufficient and -constrained situations.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in handling a broad spectrum of tasks (Lu et al., 2023; Gao et al., 2023; Kojima et al., 2022; Liu et al., 2022; Wei et al., 2024), even surpassing human performance in certain scenarios (Touvron et al., 2023; Achiam et al., 2023; Fei et al., 2022). However, in the pursuit of general artificial intelligence (AGI), LLMs are troubled by an intriguing phenomenon known as the "reversal curse" (Berglund et al., 2023; Allen-Zhu and Li, 2023; Ma et al., 2023; Zhu et al., 2024). This phenomenon occurs when a model is trained on knowledge of the form "A is B" struggles to infer the reverse relationship "B is A". For instance, a model trained on "A's parent is B" may not generalize to "B's child is A". To develop LLMs into comprehensive AGI, it is crucial to explore effective solutions to this issue.

Recently, some preliminary explorations have been made to mitigate the reversal curse. Lv et al. (2023) advocates for a shift from unidirectional to fully bidirectional attention to capture richer contextual information. Despite slight improvement, this approach suffers from a discrepancy between pre-training and fine-tuning. Golovneva et al. (2024) and Guo et al. (2024) suggest aggressive sample permutation to improve antecedent prediction. However, these methods face the problems of semantic destruction or substantial computational resources. Moreover, all of them pay little attention to delving into the underlying reasons for this phenomenon, resulting in only partial mitigation.

Intrigued by this phenomenon, our work is initiated with probing into ***the underlying reasons for the reversal curse***. Drawing inspiration from previous researchers (Li and Lewandowsky, 1995; Bireta et al., 2010; Thomas et al., 2003) who deemed that humans are profoundly influenced by knowledge storage and reasoning capability when performing knowledge reversal, we consider the following three potential reasons: (1) knowledge clarity, which originates from exposure bias in the training corpus; (2) entity correlation modeling, influenced by the specific entity order in the sample; and (3) pairwise relationship reasoning capability, affected by the path direction of reasoning. To verify these potential reasons, a series of pilot ex-

†Corresponding author.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 7518–7530
November 12-16, 2024

periments are conducted to quantify the individual influence of these potential reasons on the reversal curse. Concretely, we first collect question-answer pairs through meticulously controlled data synthesis with well-designed templates. Then, specific reference groups and experimental groups are set up to detect the individual impact of each reason on the model's reversal capability. We find that the three reasons collectively contribute to the reversal curse to varying extents, with entity correlation modeling exerting the most significant impact, followed by pairwise relationship reasoning capability and knowledge clarity.

Motivated by the analysis of these reasons, in this paper, we propose a novel **P**airwise entity **O**rder- and **R**elationship-**E**nhanced (PORE) data strategy, which facilitates bidirectional entity correlation modeling and pairwise relationship reasoning to overcome the reversal curse. Specifically, PORE reorders the entities in the given sample to build entity order-reversal and semantically preserved question-answer pairs. These pairs are exploited to augment the samples, achieving pairwise entity order and enhancing the encoding of entity correlations in both directions. PORE also employs entity-interleaved pairwise relationship data (e.g., "A is B", "D is C") , which elevates the model's capability for relationship reasoning and implicitly captures the reciprocity of relationships. Additionally, to improve the recall of reverse relationships, we highlight leveraging knowledge clarity to construct high-clarity data for PORE.

To validate the effectiveness and generalization of our method, we assemble two new datasets encompassing prevalent relationship types (author-work, company-CEO) for evaluation. Compared to previous works, our method exhibits better performance in both data-sufficient and data-constrained situations. It overcomes the reversal curse in the fine-tuning stage by Low-Rank Adaptation (LoRA) (Hu et al., 2021), which is lightweight to strengthen the reversal capability for LLMs. In summary, our main contributions include:

1) Inspired by human knowledge reversal, we delve into and quantify the individual influence of three potential reasons on the reversal curse: knowledge clarity, entity correlation modeling, and pairwise relationship reasoning capability.

2) We propose a novel data strategy, dubbed PORE, which facilitates bidirectional entity correlation modeling and pairwise relationship reasoning to overcome the reversal curse.

3) We point out how to leverage knowledge clarity to construct high-clarity data for PORE, improving the recall of reverse relationships.

4) We assemble two new datasets encompassing prevalent relationship types (author-work, company-CEO) for evaluation, facilitating future research on the reversal curse.

2 Potential Reasons for Reversal Curse

In this section, we first introduce three potential reasons for the reversal curse. Then, a series of pilot experiments are conducted to quantify the individual influence of these reasons on the reversal curse. Finally, we conclude the quantitative results and perform a subsequent qualitative analysis.

2.1 Three Potential Reasons

In contrast to previous works that pay little attention to delving into the underlying reasons for the reversal curse, we investigate it more profoundly from the perspective of human knowledge reversal. Particularly, when humans answer the reversal question "Who is B's child?" by reversing the stored forward knowledge "A's parent is B", the process can be intuitively divided into three steps (Li and Lewandowsky, 1995; Bireta et al., 2010; Thomas et al., 2003): (1) recall the knowledge related to the question (i.e., "A's parent is B"); (2) perform pairwise relationship reasoning from parent to child; and (3) model $P(A|B)$ to answer the question. Inspired by this process, we introduce three potential reasons for the reversal curse: (1) knowledge clarity, which reflects knowledge recall; (2) entity correlation modeling (e.g., $P(A|B)$, $P(B|A)$); and (3) pairwise relationship reasoning capability, namely child to parent and parent to child. We then conduct a series of pilot experiments to verify these reasons.

2.2 Pilot Experiments

Data. The 1,513 items of the relationship between the actual celebrities and their parents serve as the basic data (Berglund et al., 2023). Later, we convert this basic data into question-answer pairs pertaining to celebrity relationship to explore the knowledge storage and extraction of LLMs (Allen-Zhu and Li, 2023; Zhu and Li, 2023; Jiang et al., 2024). Specifically, for the pilot experiment on knowledge clarity, we further split the basic data into two groups via few-shot question&answer. One group consists of high knowledge clarity data, where the model can correctly recall the relevant knowledge

Potential Factors	Template		
	Basic	Reference	Experiment
Knowledge Clarity	A's parent is B	Who is A's parent? (low)	Who is C's parent? (high)
Entity Correlation Modeling	A's parent is B	A's parent is B + A's parent is whom? B	A's parent is B + B is whose parent? A
Pairwise Relationship Reasoning Capability	A's parent is B	A's parent is B + E is whose child? F	A's parent is B + F's child is whom? E

Table 1: Templates for data synthesis used in the pilot experiments. "Parent" denotes "father" or "mother". "A-F" denote specific celebrities. Low and high denote the grade of knowledge clarity.

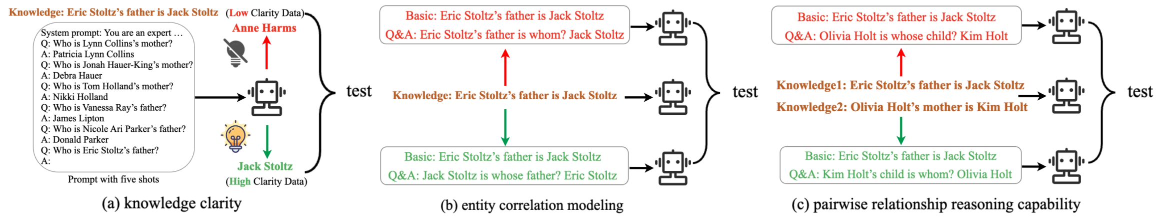

Figure 1: An illustration of the three pilot experiments on three potential reasons for the reversal curse. The red and green texts denote the data from the reference group and the experimental group. All model are evaluated on two types of reversal questions (e.g., R1: "Whose father is Jack Stoltz?" and R2: "Who is Jack Stolz's child?").

to answer the forward question. The remaining data forms the low knowledge clarity group.

Then, as shown in Table 1, specific templates are designed to meticulously control the data synthesis of the reference group and the experimental group for each potential reason. Particularly, we set the difference between the reference group and the experimental group for entity correlation modeling as "A's parent is whom? B" (i.e., $P(B|A)$) and "B is whose parent? A" (i.e., $P(A|B)$). We also set the difference in pairwise relationship reasoning capability between the reference group and the experimental group as E is whose child? F" (i.e., child to parent) and "F's child is whom? E" (i.e., parent to child). By doing so, the individual influence of these two potential reasons on the reversal curse can be clearly detected. More detailed information about the data used in the pilot experiments is provided in Appendix A.

Experimental Setup. As depicted in Figure 1, to quantify the individual influence of various potential reasons, we conduct three independent experiments. For knowledge clarity, we evaluate the pre-trained model's performance in answering questions in both the reference group and the experimental group. For entity correlation modeling and pairwise relationship reasoning capability, we first fine-tune the pre-trained model separately on the synthetic data of the reference group and the experimental group. Then, we assess the influence of these two reasons on the reversal curse by comparing the results from both groups. At the inference stage, we design two types of forward and reversal questions to reduce accidental factors, with each question preceded by five demonstrations as displayed in Figure 1a. Exact-match is utilized to determine the accuracy of generated answers. The experimental details are consistent with the main experiments, which are discussed in Section 4.3.

Quantitative Results. The results of the pilot experiments are presented in Table 2. **For knowledge clarity**, we can observe that: (i) the performance gap on the forward questions highlights the difference in knowledge clarity between the two groups of data. (ii) the experimental group's results (23.54, 23.87) are superior to the reference group's (8.21, 7.33) on the reversal questions, indicating the influence of knowledge clarity on the reversal course.

Regarding entity correlation modeling and pairwise relationship reasoning capability, it is noteworthy that all groups achieve strong performance on the forward questions. This ensures that the results of the reversal questions are minimally affected by knowledge clarity. Regarding the outcomes of these two reasons in the reversal questions, we can discern the followings: **With respect to entity correlation modeling**, (i) the results of the reference group (12.95, 10.96) and the basic group (13.21, 11.62) are on par with each other, which clearly proves that the format of the question-answer itself can not improve model's reversal capability. **In terms of pairwise relationship reasoning capability**, (ii) the performance of the reference group (11.76, 10.83) is comparable to that of the basic group (13.21, 11.62), which demonstrates that simply providing the relational word (i.e., child) can not promote the model to overcome the reversal curse. (iii) Instead, when provided with samples in-

Knowledge Clarity	F1 (Who is A's parent?) ↑	F2 (Whose child is A?) ↑	R1 (Whose parent is B?) ↑	R2 (Who is B's child?) ↑
Ref (low)	0	11.38	8.21	7.33
Exp (high)	100	50.08	**23.54**	**23.87**
Entity Correlation Modeling	F1 (Who is A's parent?) ↑	F2 (Whose child is A?) ↑	R1 (Whose parent is B?) ↑	R2 (Who is B's child?) ↑
Bas (A's parent is B)	98.68	90.36	13.21	11.62
Ref (A's parent is B + A's parent is whom? B)	98.55	91.94	12.95	10.96
Exp (A's parent is B + B is whose parent? A)	98.21	90.68	**95.37**	**95.50**
Pairwise Relationship Reasoning Capability	F1 (Who is A's parent?) ↑	F2 (Whose child is A?) ↑	R1 (Whose parent is B?) ↑	R2 (Who is B's child?) ↑
Bas (A's parent is B)	98.68	90.36	13.21	11.62
Ref (A's parent is B + E is whose child? F)	95.90	96.43	11.76	10.83
Exp (A's parent is B + F's child is whom? E)	98.94	84.28	**29.85**	**29.59**

Table 2: The pilot experimental results. Bas, Ref, and Exp denote the basic, reference, and experimental groups, respectively. F1 and F2 are two types of forward questions, while R1 and R2 are two types of reversal questions.

volving reversal relationship reasoning in the experimental groups, the performance improves (29.85, 29.59), reflecting the influence of the reversal relationship reasoning capability on the reversal curse.

2.3 Analysis on the Potential Reasons

With the quantitative results displayed in Table 2, we can conclude that the three reasons collectively contribute to the reversal curse to varying extents, with entity correlation modeling exerting the most significant impact, followed by pairwise relationship reasoning capability and knowledge clarity. Then, we qualitatively analyze how these reasons facilitate the reversal curse. For knowledge clarity, similar to humans, the deeper the knowledge is memorized, the easier it is to recall for reverse reasoning. Regarding entity correlation modeling and pairwise relationship reasoning capability, when LLMs are trained on knowledge of the form "A is B", the one-way modeling characteristic unintentionally causes the modeling of P(B|A) to be far better than P(A|B), forming asymmetric entity correlation modeling. In addition, it probably makes the model focus on reasoning the relationship from A to B while underestimating reasoning the relationship from B to A, resulting in inadequate pairwise relationship reasoning capability and awareness of the reciprocity of relationships. All of them are not conducive to answering reversal questions, thus leading to the reversal curse. Motivated by the analysis of these reasons, we propose a data strategy PORE and leverage the knowledge clarity to construct data for PORE.

3 Methodology

In this section, we detail our proposed data strategy, dubbed PORE, which facilitates bidirectional entity correlation modeling and pairwise relationship reasoning to overcome the reversal curse. As shown in Figure 2a, PORE augments the original samples with entity order-reversal and semantically preserved question-answer pairs, enhancing the encoding of reverse entity correlations. Then, PORE employs the entity-interleaved pairwise relationship data, which elevates the model's capability for reversal relationship reasoning. To improve the recall of reverse relationships, we leverage knowledge clarity to construct high-clarity data for PORE, as shown in Figure 2b.

3.1 PORE Data Strategy

According to the analysis of potential reasons in the pilot experiments, we can confirm that entity correlation modeling and pairwise relationship reasoning capability play important roles in combating the reversal curse. Note that knowledge clarity is also crucial, which will be discussed later. Generally, given knowledge of the form "A's parent is B", the forward entity correlation $P(B|A)$ and forward relationship reasoning from child to parent can be effortlessly modeled due to the one-way modeling characteristic of LLMs. Therefore, the core challenge is to improve the reverse entity correlation $P(A|B)$ and reverse relationship reasoning from parent to child, while not disturbing the corresponding forward modeling. To address this challenge, we propose a data strategy (PORE) centered on pairwise entity order and pairwise relationship reasoning, as shown in Figure 2a.

Pairwise Entity Order requires that the entities in the sample appear in both forward and reverse order. Borrowing the idea that the question-answer pair is an effective format for enhancing knowledge memorization and extraction (Allen-Zhu and Li, 2023; Zhu and Li, 2023; Jiang et al., 2024), we exploit it to construct the sample with reverse entity order. Concretely, given an original sample with forward entity order (e.g., "A's parent is B"), we flip the entity order by constructing a question about B with A as the answer (e.g., "B is whose parent?

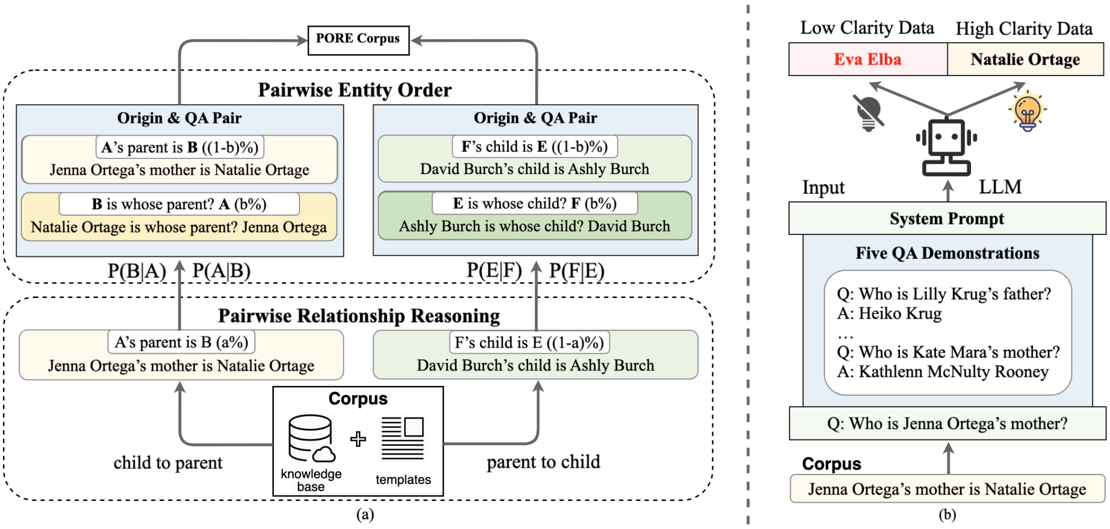

Figure 2: (a) An illustration of applying PORE to the corpus (b) An illustration of leveraging knowledge clarity to construct high-clarity data for PORE. a denotes the proportion of the entity-interleaved pairwise relationship data in the corpus. b denotes the probability of augmenting the original sample with the order-reversal and semantically preserved question-answer pair. For illustration purpose, we take an specific example from the celebrity dataset.

A"). The constructed question-answer pair then replaces the original sample with a probability of b to achieve pairwise entity order, enhancing the encoding of reverse entity correlation $P(A|B)$. Note that the semantics of the question-answer pair are consistent with the original sample, aiding in reducing disturbances to the forward modeling $P(B|A)$.

Pairwise Relationship Reasoning requires the sample to contain both forward and reverse paths of relationship reasoning. Concretely, we split the corpus into two parts with a proportion of a, where each part implements one direction of relationship reasoning, thereby achieving pairwise relationship reasoning. By doing so, the model is able to pay more attention to reasoning the reverse relationship from parent to child and capturing the reciprocity of the relationship. It should be noted that the entities are interleaved (e.g., "A's parent is B, F's child is E"), which will not affect reverse entity correlation modeling $P(A|B)$.

To jointly facilitate reverse entity correlation and reverse relationship reasoning, we first partition the corpus into two parts, with a proportion of a. Each part implements relationship reasoning in one direction. For each original sample in both parts, we augment it with the entity order-reversal and semantically preserved question-answer pair with a probability of b to obtain the PORE corpus. Through varying the values of a and b, we can adjust the weight of forward and reverse modeling. Finally, given the sample $Y = y_1, ..., y_s$ in the PORE corpus, the LLM is trained by minimizing the negative log-likelihood loss L_θ as below:

$$L_\theta = -\sum_{k=1}^{s} logp(y_k|y_{<k}, \theta). \qquad (1)$$

where θ denotes the trainable parameters of the LLM. It should be noted that we replace the samples with a certain probability b, which will not increase the overall number of training samples.

3.2 Data Construction with Knowledge Clarity

Knowledge clarity generally originates from exposure bias in the pre-training corpus. Intuitively, the more frequently the knowledge appears in the corpus, the higher its clarity, as the LLM has more opportunities to memorize it. Considering the massive scale and inaccessibility of the pre-training corpus, it is nearly impossible to mitigate this bias by ensuring all knowledge has high-clarity to alleviate the reversal curse. This situation is akin to the difficulty humans face in thoroughly remembering all knowledge. Therefore, we consider another perspective: leveraging knowledge clarity to construct data for PORE to handle data constraints. The specific details are shown in Figure 2b. Concretely, we first design a system prompt to guide the model in performing question&answer. For each example, we provide an additional five demonstrations to improve the model's knowledge recall capability. If the model can recall the relevant knowledge to answer the question, the corresponding sample is regarded as high-clarity knowledge. Finally, we apply PORE to these samples, as we believe that the model can harness the high-clarity data to improve the recall of the reversal relationship, contributing to mitigating the reversal curse. This is analogous to the human process of recalling familiar knowledge (high-clarity) to infer the answer. Combining knowledge clarity with PORE, our method is expected to combat the reversal curse in both data-sufficient and data-constrained situations.

4 Experiments

In this section, we introduce the datasets, baselines, and experimental settings. Then, we present experimental results and provide a detailed analysis.

4.1 Datasets

To verify the effectiveness and generalization of our method, we conduct experiments on an available celebrity relationship dataset (Berglund et al., 2023) used in the pilot experiments, and two newly assembled datasets containing 2,000 items of the Author-Work relationship and 1,697 items of the Company-CEO relationship. For each sample in these datasets, we tailor two types of forward and reversal questions for evaluation. We ensure that the model can not encounter the same questions during training. Further details about the datasets can be found in Appendix A.

4.2 Baselines

(1) Llama (Touvron et al., 2023), **(2) GPT-3.5**, and **(3) GPT-4** (Achiam et al., 2023) are powerful large language models. We directly evaluate them with few-shot question&answer to expose the reversal curse. **(4) Llama (SFT)** is first fine-tuned on the dataset and then evaluated its capability to answer reversal questions. **(5) Reverse** adopts a word-level reverse training method to model the right-to-left word correlation, reducing the effect of the reversal curse. **(6) BICO** (Lv et al., 2023) modifies unidirectional attention to fully bidirectional attention to capture richer contextual information for knowledge reversal. **(7) SPT** (Guo et al., 2024) utilizes semantic permutation training to improve antecedent prediction. **(8) RSP** (Golovneva et al., 2024) leverages segment permutation training to sneakily learn the knowledge in its reverse direction. We choose its three variants according to the maximum length of segmentation (i.e., $k = 2, 3, 5$) for a fair comparison. More details about the baseline are shown in Appendix B.

4.3 Experimental Settings

We compare our method PORE with existing methods under two experimental settings. One is the data-sufficient situation, which uses the full dataset for training and evaluates on the tailored questions. The other is a more challenging setting that considers the realistic scenario where we can only apply the data strategy to limited data. In this case, we use partial data for training and then evaluate on the tailored questions of the remaining data. Note that the training data is divided by knowledge clarity. We engage Llama2-7b (Touvron et al., 2023) as the backbone, with only 4.5% of the parameters fine-tuned thanks to LoRA (r=128). During training, the hyper-parameters are set as follows: learning rate: 2e-5, batch size: 10, and epoch: 16 for all datasets. For inference, following previous works (Berglund et al., 2023; Guo et al., 2024), exact-match is used to determine the accuracy of generated answers. More details on the experimental settings can be found in Appendix C.

4.4 Experimental Results

The evaluation results on the three datasets under the data-sufficient situation are reported in Table 3. We can find that PORE outperforms other baselines in almost all metrics for both types of questions (forward or reversal), indicating its powerful efficacy in combating the reversal curse while preserving the capability to answer the forward questions.

Further, we can observe that: (i) the performance degradation on reversal questions in powerful LLMs (e.g., Llama, GPT3.5) reveals the reversal curse. This also illustrates that the curse can not disappear with an increase in model size (GPT-4) or simple fine-tuning. (ii) Bidirectional attention (BICO) and word-level reverse training show improvement on reversal questions, proving the benefits of reverse entity correlations. However, the effect is limited because both methods disrupt the integrity of the entity, resulting in inadequate entity correlation modeling. This problem is highlighted in the dataset (e.g., Author-Work) with longer entity lengths. (iii) PORE exhibits superiority over SPT and RSP, which adopt aggressive sample permutation to improve the antecedent prediction. It is probably because our method avoids semantic corruption while exploiting entity-reversal question-answer pairs to more explicitly encode reverse entity correlation. Moreover, PORE employs entity-interleaved pairwise relationship data to elevate the

Models	Celebrity				Author-Work				Company-CEO			
	F1↑	F2↑	R1↑	R2↑	F1↑	F2↑	R1↑	R2↑	F1↑	F2↑	R1↑	R2↑
Llama	40.29	27.61	15.72	13.74	16.80	17.60	3.90	3.60	12.15	12.38	9.43	12.38
GPT-3.5	57.99	46.90	40.82	31.18	43.50	44.70	8.40	8.50	31.96	33.49	21.34	22.29
GPT-4	65.39	40.95	41.48	39.50	51.10	49.80	9.60	8.40	42.92	42.69	27.24	25.71
Llama (SFT)	98.68	90.36	13.21	11.62	<u>92.10</u>	<u>91.50</u>	5.60	5.80	**87.50**	<u>83.37</u>	16.86	18.04
Reverse	98.81	94.19	33.55	30.78	91.80	91.00	6.80	7.00	70.40	67.45	33.14	32.43
BICO	**99.60**	<u>96.43</u>	22.19	16.64	79.50	76.40	6.60	6.90	78.07	75.83	20.87	20.28
SPT	<u>99.08</u>	96.30	61.43	59.84	-	-	-	-	-	-	-	-
RSP (k=2)	59.18	62.75	65.65	65.92	23.40	24.10	13.40	13.40	33.02	34.08	39.62	36.67
RSP (k=3)	71.73	75.56	78.20	77.15	36.10	36.10	18.20	17.00	46.58	47.17	50.83	50.00
RSP (k=5)	92.47	87.98	<u>84.81</u>	<u>83.32</u>	67.60	67.50	<u>24.30</u>	<u>24.20</u>	75.24	73.82	<u>64.50</u>	<u>63.92</u>
PORE	98.81	**97.23**	**96.96**	**97.49**	**93.30**	**93.30**	**88.70**	**85.10**	<u>86.67</u>	**85.73**	**91.75**	**93.16**

Table 3: Evaluation results on three datasets under the data-sufficient situation. F1, F2 and R1, R2 are two types of forward and reversal questions. k denotes the maximum segment length. Best in bold, the second with an underline.

Models	Low clarity				High clarity				Delta
	F1↑	F2↑	R1↑	R2↑	F1↑	F2↑	R1↑	R2↑	$(\triangle_{R1} + \triangle_{R2})/2 \uparrow$
Llama	0	12.38	6.66	6.02	0	12.38	6.66	6.02	-
Llama (SFT)	98.41	82.54	9.21	6.67	98.73	85.71	9.52	6.98	0.31
Reverse	99.05	80.95	8.89	6.67	98.73	90.48	12.70	11.43	4.29
BICO	**99.37**	**89.52**	10.48	7.30	**99.37**	**96.83**	13.65	8.89	2.38
SPT	97.46	83.81	10.61	10.19	98.73	83.81	14.87	13.65	3.86
RSP (k=5)	97.14	83.49	9.25	8.62	97.78	81.59	14.29	13.97	5.19
PORE	97.14	87.94	**10.89**	**10.42**	98.10	92.38	**16.19**	**17.14**	6.01

Table 4: Evaluation results on the celebrity relationship dataset under the data-constrained situation. F1, F2 and R1, R2 denote two types of forward and reversal questions. k denotes the maximum segment length. Delta denotes the average gain in reversal performance. $\triangle_{Ri}$ denotes the gap between high-clarity data and low-clarity data in Ri.

model's capability for relationship reasoning and to capture the reciprocity of relationships.

The evaluation results on the celebrity relationship dataset under the data-constrained situation are presented in Table 4. We can discern that, regardless of the knowledge clarity of the training data, the results on the reversal questions for all models are far behind those in data-sufficient situation. It could be that the model is unable to encode the reverse correlations of entities in the test samples, resulting in poor performance. This aligns with the quantitative results of the pilot experiment, which indicate that entity correlation modeling exerts the most significant impact. Nevertheless, PORE still surpasses other baselines in answering the reversal questions. We believe this is because the entity-interleaved pairwise relationship data facilitates the model in understanding the reciprocity of relationships and enhancing its capability for reversal relationship reasoning.

Comparing the reversal capabilities of the same model trained on data with different knowledge clarity, we can find that the high-clarity data consistently outperforms the low-clarity data. We believe the reason is that the high-clarity data is easier to recall, enabling the model to exploit the implicit reversal relationship entailed in the training data to infer the reversal answers. In a word, compared to previous methods, the superior performance in the above two experiments demonstrates the effectiveness of PORE and the superiority of constructing data with knowledge clarity.

4.5 Ablation Study

To explicitly illustrate the effectiveness of our method PORE, we conduct ablation studies to validate its core design on three datasets under the data-sufficient situation. As shown in Table 5, we present the following three ablation variants: **(1) -w/o PRR** removes pairwise relationship reasoning by discarding the entity-interleaved pairwise relationship data. **(2) -w/o PEO** removes pairwise

Models	Celebrity				Author-Work				Company-Ceo			
	F1↑	F2↑	R1↑	R2↑	F1↑	F2↑	R1↑	R2↑	F1↑	F2↑	R1↑	R2
PORE	98.81	**97.23**	**96.96**	**97.49**	**93.30**	**93.30**	**88.70**	**85.10**	86.67	85.73	**91.75**	**93.16**
- w/o PRR	98.21	90.68	95.37	95.50	92.40	91.80	85.70	83.30	86.67	85.96	91.04	92.22
- w/o PEO	**99.21**	84.94	27.87	27.92	89.90	90.10	10.40	10.20	**89.27**	**86.91**	18.99	18.87
- w/o BOTH	98.68	90.36	13.21	11.62	92.10	91.50	5.60	5.80	87.50	83.37	16.86	18.04

Table 5: Ablation study results on three datasets under the data-sufficient situation. PRR and PEO denote pairwise relationship reasoning and pairwise entity order, respectively.

entity order by discarding the entity-reversal and semantically preserved question-answer pairs. **(3) -w/o BOTH)** is the combination of **(1)** and **(2)**.

Specifically, we can draw the following inferences based on the results in Table 5: (i) Removing pairwise relationship reasoning (-w/o PRR) leads to a slight performance drop, and the extent of the drop is affected by the relationship type in the dataset. We attribute this to the varying levels of mastery of different relationship reasoning during the pre-training stage. (ii) Removing pairwise entity order (-w/o PEO) results in significant performance degradation, reflecting the crucial role of entity correlation modeling in overcoming the reversal curse. (iii) Removing both pairwise relationship reasoning and pairwise entity order (-w/o BOTH) causes the worst performance, indicating the necessity of each core design. The results of the ablation studies confirm the analysis of the pilot experiments, where entity correlation modeling exerts the most significant impact, followed by pairwise relationship reasoning capability.

Methods	Data Costs	F1 ↑	F2 ↑	R1 ↑	R2 ↑
SPT	3M	**99.08**	96.30	61.43	59.84
RSP (k=5)	2M	92.47	87.98	84.81	83.32
PORE	M(1+α), $\alpha \in (0, 1)$	98.81	**97.23**	**96.96**	**97.49**

Table 6: The data costs measured by the number of times where each sample needs to be expanded.

Datasets	Training Time	Inference Time	Inference Time
Celebrity	16min 48s	2min 47s	20524 MiB
Author-Work	23min 47s	4min 11s	21338 MiB
Company-Ceo	19min 02s	3min 16s	20600 MiB

Table 7: The computational costs across all datasets in fine-tuning stage by LoRA (only 4.5% parameters fine-tuned) with only a single A100-40G GPU.

5 Efficiency Analysis

To perform a more comprehensive comparison with other baselines, we conduct an efficiency analysis, including data costs and computational costs. **For the former**, we measure them by the number of times where each sample needs to be expanded. The results are reported in Table 6. Taking a specific sample on the celebrity datasets as an example (e.g., "A's parent is B"), according to the proposed PORE, α% of samples are first used to construct entity-interleaved pairwise relationship data (i.e., "B's child is A"). Then, PORE requires each sample to construct an order-reversal and semantically question-answer pair (i.e., "B is whose parent? A" or "A is whose parent? B"). Hence, the data costs for a corpus with M samples can be formulated as:

$$2 \times M \times \alpha + M(1 - \alpha) = M(1 + \alpha), \alpha \in (0, 1). \quad (2)$$

For the latter, we present the computational costs of our proposed PORE across all datasets in the fine-tuning stage by Low-Rank Adaptation (LoRA, only 4.5% parameters fine-tuned) with only a single A100-40G GPU, are shown in Table 7. It should be noted that compared to other baselines, our method does not increase the number of training samples, because PORE replaces the original samples with constructed semantically preserved samplers with a certain probability instead of doubling the training samples.

6 Related Work

Reversal Curse in LLMs The reversal curse was first revealed by concurrent work (Berglund et al., 2023; Allen-Zhu and Li, 2023). They showed that the curse is widespread in auto-regressive large language models and can not be addressed by merely increasing the model size or simply fine-tuning. Recently, some initial explorations have been conducted to alleviate the reversal curse. Lv et al. (2023) modified unidirectional attention to fully bidirectional attention to capture richer contextual information. Golovneva et al. (2024) and Guo et al. (2024) employed aggressive sample permutation to improve antecedent prediction. However, these methods face the problem of semantic destruction or substantial computational resources. Moreover, they pay little attention to delving into the underlying reasons for the reversal curse, resulting in only partial mitigation. In this paper, we investigate and quantify the underlying reasons for the reversal curse and then propose a novel data strategy (PORE) to overcome the reversal curse.

Reversal Curse in Humans "Are humans troubled by the reversal curse? We guess the answer is yes". For instance, we always find it easier to recite phone numbers and the alphabet forward than backward, and this applies to other serialized memories as well (Berglund et al., 2023; Guitard et al., 2020). Moreover, previous works have claimed that the forward and backward memory mechanisms are different (Bireta et al., 2010; van Kerkoerle et al., 2023), and recalling backward memories is harder than forward memory (Li and Lewandowsky, 1995; St Clair-Thompson and Allen, 2013; Guitard et al., 2020; Geva et al., 2020). Thomas et al. (2003) pointed out that backward recall involves repeated covert cycles of forward recall. There is currently no clear research exploring the connection between the reversal curse of humans and the reversal curse of LLMs. In this paper, inspired by the specific human cognition aspects of "forward/backward memory recall (Bireta et al., 2010; van Kerkoerle et al., 2023)" and "working memory mechanism (Thomas et al., 2003; Guitard et al., 2020)" in human knowledge reversal, we propose a specific method, PORE, to combat the reversal curse.

7 Conclusions

In this paper, we delve into and quantify the individual influence of three potential reasons on the reversal curse: 1) knowledge clarity 2) entity correlation modeling, and 3) pairwise relationship reasoning capability. We find that these reasons collectively contribute to the reversal curse to varying extents, with entity correlation modeling exerting the most significant impact, followed by pairwise relationship reasoning capability and knowledge clarity. Motivated by the analysis of these reasons, we propose a novel PORE data strategy, which facilitates bidirectional entity correlation modeling and pairwise relationship reasoning to overcome the reversal curse. We also point out how to leverage knowledge clarity to construct data for PORE. Extensive experimental results on available and two newly assembled datasets demonstrate the effectiveness and generalization of our method in both data-sufficient and data-constrained situations. In the future, we will explore PORE for automated data synthesis and probe its scalability and practicality in the massive-scale pre-training stage. We will also attempt other reasoning format, such as change reasoning (Lu et al., 2024).

Limitations

Despite the impressive results of our method, we have to admit our work has the following limitations: (1) to meticulously quantify the influence of potential reasons and conduct a fair comparison across various baselines, our work adopts strict templates to control data synthesis, which can be laborious. With the disclosed insights in this paper, in real-world applications, we plan to directly utilize existing ChatLLMs (e.g., ChatGPT) to complete automatic data synthesis without relying on complex template design. (2) We validate the effectiveness of our method on a single dataset in the fine-tuning stage with LoRA and have not yet explored its scalability and practicality in the pre-training stage with a massive-scale pre-training corpus, which is also our future work.

Acknowledgements

This research was funded by the National Natural Science Foundation of China (62206267). We sincerely thank Wei Jia, Nayu Liu for their constructive collaboration.

References

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

Zeyuan Allen-Zhu and Yuanzhi Li. 2023. 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*.

Tamra J Bireta, Sheena E Fry, Annie Jalbert, Ian Neath, Aimée M Surprenant, Gerald Tehan, and Georgina Anne Tolan. 2010. Backward recall and benchmark effects of working memory. *Memory & cognition*, 38:279–291.

Zhaoye Fei, Yu Tian, Yongkang Wu, Xinyu Zhang, Yutao Zhu, Zheng Liu, Jiawen Wu, Dejiang Kong, Ruofei Lai, Zhao Cao, et al. 2022. Coarse-to-fine: Hierarchical multi-task learning for natural language understanding. *arXiv preprint arXiv:2208.09129*.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Pal: Program-aided language models. In *International Conference on Machine Learning*, pages 10764–10799. PMLR.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2020. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913*.

Olga Golovneva, Zeyuan Allen-Zhu, Jason Weston, and Sainbayar Sukhbaatar. 2024. Reverse training to nurse the reversal curse. *arXiv preprint arXiv:2403.13799*.

Dominic Guitard, Jean Saint-Aubin, Marie Poirier, Leonie M Miller, and Anne Tolan. 2020. Forward and backward recall: Different visuospatial processes when you know what's coming. *Memory & Cognition*, 48:111–126.

Qingyan Guo, Rui Wang, Junliang Guo, Xu Tan, Jiang Bian, and Yujiu Yang. 2024. Mitigating reversal curse via semantic-aware permutation training. *arXiv preprint arXiv:2403.00758*.

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

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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

Shu-Chen Li and Stephan Lewandowsky. 1995. Forward and backward recall: Different retrieval processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(4):837.

Nayu Liu, Kaiwen Wei, Xian Sun, Hongfeng Yu, Fanglong Yao, Li Jin, Guo Zhi, and Guangluan Xu. 2022. Assist non-native viewers: Multimodal cross-lingual summarization for how2 videos. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6959–6969.

Zhicong Lu, Li Jin, Ziwei Chen, Changyuan Tian, Xian Sun, Xiaoyu Li, Yi Zhang, Qi Li, and Guangluan Xu. 2024. Relation-aware multi-pass comparison deconfounded network for change captioning. *IEEE Transactions on Circuits and Systems for Video Technology*.

Zhicong Lu, Li Jin, Guangluan Xu, Linmei Hu, Nayu Liu, Xiaoyu Li, Xian Sun, Zequn Zhang, and Kaiwen Wei. 2023. Narrative order aware story generation via bidirectional pretraining model with optimal transport reward. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6274–6287.

Ang Lv, Kaiyi Zhang, Shufang Xie, Quan Tu, Yuhan Chen, Ji-Rong Wen, and Rui Yan. 2023. Are we falling in a middle-intelligence trap? an analysis and mitigation of the reversal curse. *arXiv preprint arXiv:2311.07468*.

Jun-Yu Ma, Jia-Chen Gu, Zhen-Hua Ling, Quan Liu, and Cong Liu. 2023. Untying the reversal curse via bidirectional language model editing. *arXiv preprint arXiv:2310.10322*.

Helen L St Clair-Thompson and Richard J Allen. 2013. Are forward and backward recall the same? a dual-task study of digit recall. *Memory & cognition*, 41:519–532.

John G Thomas, Haley R Milner, and Karl F Haberlandt. 2003. Forward and backward recall: Different response time patterns, same retrieval order. *Psychological Science*, 14(2):169–174.

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

Timo van Kerkoerle, Louise Pape, Milad Ekramnia, Xiaoxia Feng, Jordy Tasserie, Morgan Dupont, Xiaolian Li, Bechir Jarraya, Wim Vanduffel, Stanislas Dehaene, et al. 2023. Brain mechanisms of reversible symbolic reference: a potential singularity of the human brain. *bioRxiv*, pages 2023–03.

Kaiwen Wei, Jingyuan Zhang, Hongzhi Zhang, Fuzheng Zhang, Di Zhang, Li Jin, and Yue Yu. 2024. Chain-of-specificity: An iteratively refining method for eliciting knowledge from large language models. *arXiv preprint arXiv:2402.15526*.

Hanlin Zhu, Baihe Huang, Shaolun Zhang, Michael Jordan, Jiantao Jiao, Yuandong Tian, and Stuart Russell. 2024. Towards a theoretical understanding of the'reversal curse'via training dynamics. *arXiv preprint arXiv:2405.04669*.

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

A Details on Datasets

Celebrity Relationship Dataset: The celebrity relationship (Bireta et al., 2010) dataset collected a list of the top 1000 most popular celebrities from IMDB and queries GPT-4 (Achiam et al., 2023) for their parents. Consequently, it contained 1,513 items of relationships between actual celebrities and their parents. The relational words for a parent-child pair are "father/mother" and "child".

Author-Work Dataset: The author-work dataset is derived from the DBLP (Digital Bibliography & Library Project) bibliography [*], a widely used and publicly available bibliographic database of computer science research papers and proceedings. We randomly selected 2,000 author-work pairs from it and limited each author to no more than three books. The relational words for an author-work pair are "book" and "work".

Company-Ceo Dataset: The company-ceo dataset is crawled from Forbes.com [†]. We first selected the data on the top 2000 companies and their corresponding chief executive officers. Then, we filtered out the N/A entries and consequently obtained 1,697 items of relationships between actual companies and their corresponding chief executive officers. The relational words for a company-CEO pair are "company" and "CEO".

Regarding the dataset in the pilot experiments, we serve the celebrity relationship dataset as the raw data and further process it to obtain the basic data as follows: given the parent-child pair, (i) we first use it to fill the well-designed templates as shown in 1. Then, (ii) we combine these data to construct the control and the experimental groups for different potential reasons. Taking a specific sample in the pilot experiment on entity correlation modeling as an example, if the parent-child pair is (Eric Stoltz, Jack Stoltz, father), we can obtain the data in the reference group as the original sample "Eric Stoltz 's father is Jack Stoltz" and as the Q&A sample "Eric Stoltz's father is whom? Jack Stoltz". We can also obtain the data in the experimental group as the original sample "Eric Stoltz 's father is Jack Stoltz" and as the Q&A sample as "Jack Stoltz is whose father? Eric Stoltz". Note that during actual training, the original sample and the Q&A sample will not appear in an epoch at the same time. We adopt the sampling technique with the probability of b to keep the total number of training samples unchanged.

Last, (iii) we construct two forward and two reversal questions as the test samples for each pair. For example, the two forward questions for the pair (Eric Stoltz, Jack Stoltz, father) are "Who is Eric Stoltz's parent?" and "Whose child is Eric Stoltz?". We can also obtain the two reversal questions "Whose parent is Jack Stoltz?" and "Who is Jack Stoltz's child?". It should be noted that for each type of question, we adopt different relational words to formulate the questions, reducing the accidental factors. We ensure that the model will not encounter the same questions during training. Due to the use of entity-interleaved data in the pilot experiment on pairwise relationship reasoning capability, we can only obtain half of the full data to construct the test data. We also ensure that the test samples are consistent across all pilot experiments.

Regarding the dataset in the main experiments, the overall data processing is the same as in the pilot experiments. The difference lies in the templates as follows:

(1) For the celebrity dataset, the template for the original sample is "A's parent is B". The template for the Q&A sample is "B is whose parent? A". The template for the entity-interleaved sample is "F's child is E". The template for the Q&A sample of entity-interleaved sample is "E is whose child? F". The templates for two forward questions are "Who is A's parent?" and "Whose child is A", the templates for two reversal questions are "Whose parent is B?" and "Who is B's child?"

(2) For the author-work dataset, the template for the original sample is "A's author is B". The template for the Q&A sample is "B is author of what? A". The template for the entity-interleaved sample is "F's work is E". The template for the Q&A sample of the entity-interleaved sample is "E's work is what? F". The templates for two forward questions are "Who is A's author?" and "Whose work is A", the templates for two reversal questions are "Whose is B author of?" and "Who is B's work?"

(3) For the company-ceo dataset, the template for the original sample is "A's company is B". The template for the Q&A sample is "Whose company is B? A". The template for the entity-interleaved sample is "F's CEO is E". The template for the Q&A sample of entity-interleaved sample is "E is CEO of what? F". The templates for two forward questions are "What is A's company?" and "What is A CEO of?", the templates for two reversal ques-

[*] dblp.uni-trier.de/

[†] https://www.forbes.com/lists/global2000/

tions are "Whose company is B?" and "Who is B's CEO?"

In a word, in this paper, to precisely find the underlying reasons for the reversal curse, we carefully design templates to control the data synthesis in the pilot experiments. To ensure a fair comparison across all baselines, we also utilize templates to construct data in the main experiments. In fact, we could leverage LLMs to automatically generate the above data in real scenarios. The above dataset will be available in https://github.com/lzc-nazarite/PORE

B Details on Baselines

The baselines in this work can be intuitively divided into three categories:

(1) Few-shot LLMs: This denotes LLama2-7b, GPT-3.5, and GPT-4. We directly evaluate them with five-shot question&answer to expose the reversal curse. The five demonstrations are randomly selected from the training datasets. The difference in evaluation on different datasets lies in the system prompt.

For the celebrity dataset, the system prompt is: "You are an expert when it comes to celebrities from various fields, such as actors, singers, and producers, and their family relations. You answer questions concisely, with only the specific answer or 'I don't know'".

For the author-work dataset, the system prompt is: "You are an expert when it comes to books from various fields, such as science, literature, and technology, and their author relationships. You answer questions concisely, with only the specific answer or 'I don't know'".

For the company-ceo dataset, the system prompt is "You are an expert when it comes to companies from various fields, such as banking and financial services, technology, oil and gas, and their chief executive officer (CEO) relationships. You answer questions concisely, with only the specific answer or 'I don't know'". A specific example of the few-shot question&answer on the celebrity dataset is shown in Figure 3.

(2) BICO and Reverse Training: BICO (Lv et al., 2023) modifies unidirectional attention to fully bidirectional attention to capture richer contextual information. Reverse training (Guo et al., 2024; Golovneva et al., 2024) flips the sentence in its reverse direction and then feeds to the LLMs.

(3) Sampler Permutation Training Methods:

System prompt: You are an expert when it comes to celebrities from various fields, such as actors, singles, and producers, and their family relations. You answer questions concisely, with only the specific answer or 'I don't know'

Q: Who is Lynn Collins's mother?
A: Patricia Lynn Collins
Q: Who is Jonah Hauer-King's mother?
A: Debra Hauer
Q: Who is Tom Holland's mother?
A: Nikki Holland
Q: Who is Vanessa Ray's father?
A: James Lipton
Q: Who is Nicole Ari Parker's father?
A: Donald Parker
Q: Who is Eric Stoltz's father?
A:

Figure 3: An specific example of few-shot evaluation on the celebrity relationship dataset

These methods adopt aggressive sample permutation improve antecedent prediction. Golovneva et al. (2024) heuristically choose maximum segment length k and then split the sentence into random segments, where each segments length is shorter than k. Last, it randomly permutes all segments to feed into the LLM. Guo et al. (2024) first utilize LLMs to segment sentences into semantic units. Then it randomly permutes all units to feed into the LLM.

C Implementation Details

The Llama2-7b is engaged as the backbone and merely 4.5% parameters are fine-tuned thanks to LoRA (Hu et al., 2021)(r=128). During training, the hyper-parameters are learning rate: 2e-5, batch size: 10, and epoch: 16 for all datasets. The Adam (Kingma and Ba, 2014) optimizer with β_1=0.9, β_2=0.999 is leveraged to optimize the model by minimizing the loss in (1). The Linear learning rate scheduler is also implemented with a warmup ratio of 0.03. The proportion a and the probability b are set to 0.5 to balance the forward and reverse modeling. For each sample in inference, we employ the checkpoint from the last epoch to evaluate the model's performance on the tailored questions. Following the previous works (Berglund et al., 2023; Guo et al., 2024), the exact-match is utilized to determine the accuracy of generated answers. Both training and inference are implemented on a single A100-40G GPU.

D Concerns on Prompt Variation

To ensure a fair comparison with the previous researchers (Guo et al., 2024; Golovneva et al., 2024), we only consider two sets of paraphrased questions consistent with their works to evaluate

Paraphrased questions	SPT (F) ↑	SPT (R) ↑	RSP (k=5,F) ↑	RSP (k=5,R) ↑	PORE (F) ↑	PORE (R) ↑
1) F: Who is A's parent? R: Whose parent is B?	**99.08**	61.43	92.47	84.81	98.91	**96.96**
2) F: Whose child is A? R: Who is B's child?	96.30	59.84	87.98	83.32	**97.23**	**97.49**
3) F: A's parent is whom? R: B's child is whom?	**99.34**	60.77	92.21	82.43	97.89	**97.36**
4) F: Who is the parent of A? R: B is the parent of whom?	96.34	60.11	93.00	82.17	**98.41**	**97.23**
5) F: A is the parent of whom? R: Who is the child of B?	96.83	61.82	89.70	84.15	**98.28**	**96.83**
6) F: Who is A's legal guardian? R: Whose legal guardian is B?	96.57	62.48	87.45	82.03	**97.36**	**96.30**
7) F: Who is the legal guardian of A? R: B is the legal guardian of whom?	96.57	59.97	86.92	82.69	**96.96**	**96.83**
8) F: A is whose offspring? R: B's offspring is whom?	94.19	60.77	85.60	82.69	**97.75**	**98.28**
9) F: A is the offspring of whom? R: Who is the offspring of B?	96.83	61.56	89.43	84.28	**98.02**	**97.36**
10) Average Performance	96.89	60.97	89.42	83.16	**97.86**	**97.18**

Table 8: The performance on the celebrity dataset across various prompts.

the forward and reversal performance of LLMs. **To minimize the concern that the validity of the disclosed insights might change when the prompting variations are introduced, we add the number of sets of paraphrased questions on the celebrity dataset to nine, demonstrating the effectiveness of our method under different paraphrased questions.** The results are shown in the following Table (F denotes the forward performance, R denotes the reversal performance). We acknowledge that it is hard to completely eliminate this concern, because it is impossible to attempt all types of paraphrased questions. We hope that the consistently superior performance under varied paraphrased questions can further prove the validity of the disclosed insights as much as possible.

E Details on Knowledge Clarity

The core insight of dividing data into groups with different knowledge clarity based on the model's answer is to reveal how well this model has grasped different knowledge in the data. Then, PORE is performed on the divided high-clarity data that depends on the results of divided results to improve this model's performance. In order to make the model harness high-clarity knowledge which is based on its own mastery of different knowledge in the data to improve the recall of the reversal relationship, the model for the divided data and fine-tuned on the high-clarity data should remain the same. Therefore, for a specific model (e.g., Llama2-7b), there is no need to pay attention to the different divisions of knowledge clarity in other models, just adopting its own division.