Investigating Reversal Generalization in LLMs via Data Augmentation & Contrastive Loss

Julia Kashimura and **Laya Reddy** and **Tara Shukla**

Abstract

Autoregressive large language models (LLMs) trained on directional fact pairs ("A is B") exhibit a severe ordering bias: they answer "Who is A?" accurately but fail on the equivalent "What is B's name?" queries, a phenomenon called the *Reversal Curse*. In this paper, we (1) reproduce the reversal failure on synthetic "name→ description" and "description → name" benchmarks; (2) design two data-augmentation pipelines, reasoning-chain paraphrases and negation-based contrastive examples, to enrich the fine-tuning corpus; and (3) introduce a *symmetric contrastive loss* in our fine-tuning objective that aligns embeddings of each fact pair ("A is B" ⇔ "B is A"). Our results show moderate gains in bidirectional knowledge after fine-tuning with augmented prompting, but a reduction in performance with contrastive loss; overall, we find that data augmentation is a valid strategy for enhancing two-way knowledge acquisition and mitigating the Reversal Curse in LLMs.

1 Introduction

Large language models learn factual associations in one direction (e.g. "The composer of 'Abyssal Melodies' is Uriah Hawthorne"). When asked the same-order query ("Who composed 'Abyssal Melodies'?"), they perform reliably; however, the inverted query ("Who is Uriah Hawthorne?") often fails. This phenomenon is called the *Reversal Curse*.

To address this, we first reproduce the ordering failure on a synthetic benchmark of 30 celebrity–fact pairs, each paraphrased and split into "name→description" and "description→name" fine-tuning sets. Despite near-perfect accuracy on forward queries, reverse-query performance remains at chance.

We then propose two complementary solutions. First, we augment the standard cross-entropy objective with a *symmetric contrastive loss* that brings embedding representations of each fact and its inversion closer together during fine-tuning. Second, we create two data-augmentation pipelines: (a) *reasoning-chain paraphrases* that rephrase facts with short causal or relational chains in both directions, and (b) *negation-based contrastive examples* that teach the model to reject distractor mappings via explicit "X is **not** Y" prompts.

Through controlled ablations, we demonstrate that (i) the contrastive loss alone significantly improves reversed-query accuracy, (ii) each augmentation strategy yields further gains, and (iii) combining loss and augmentations produces the strongest bidirectional generalization—all while preserving forward-query performance. The remainder of this paper is organized as follows: Section 5 details our replication of the Reversal Curse; Section 6 describes the data-augmentation pipelines; Section 7 presents the symmetric contrastive fine-tuning objective; and Section 7 concludes with limitations and future directions.

2 Related Work

2.1 LLM Knowledge Generalization

It is a known fact that many LLMs struggle to generalize bidirectional and hierarchical knowledge. One of these limitations is the aforementioned Reversal Curse, explored in depth in (Berglund et al., 2024), which we will use as the foundation for our paper. In this paper, the authors demonstrate that LLMs will not generalize knowledge to the reverse direction in which they learned it: their approach includes finetuning experiments on synthetic data, and testing the model's ability to answer prompts in both the original order (using the name to retrieve the description) and the reversed order (using the description to retrieve the name). Results showed that models achieved high accuracy when the question order matched the finetuning data order but performed at or near chance level (accuracy close

to 0%) when the order was reversed. Other work on LLM knowledge acquisition and generalization includes (Cheang et al., 2023). The authors investigate a related limitation: the temporal generalization abilities of pre-trained LLMs. They explore how models struggle to generalize to future data that contains knowledge conflicts with their parametric memory; findings reveal that models often generate hallucinations containing outdated information when processing documents from temporal periods beyond their training data, particularly when entities have evolved relationships over time. Additional research has explored how LLMs handle structured knowledge representation. Work by (Xu et al., 2024) demonstrates that LLMs can fail to maintain consistency across different representations of the same facts. This limitation extends beyond temporal constraints to encompass other transformations, such as syntactic variations and logical equivalences. Overall, past work underscores the fact that LLMs exhibit substantial limitations in flexible knowledge generalization across different dimensions; our paper implements several solutions to address these issues.

2.2 Data Augmentation

Data augmentation for LLMs has received increasing attention as a method to improve model robustness, generalization, and bidirectional understanding. We review relevant work on data augmentation techniques, especially those could enhance bidirectional knowledge reversal capabilities. Firstly, in (Dhole et al., 2021), the authors present NL-Augmenter, a natural language data augmenter that creates data transformations such as rephrasing, causal negation, number to word, or multilinguality to text; they prove that these augmentations improve the robustness of a suite of NLP models in text classification and comparison tasks. In (Kaushik et al., 2019), Kaushik, et.al focus on data augmentation for classifier and natural language inference (NLU) tasks; they utilize counterfactual labels to prove that fine-tuning with two-way counterfactual data will improve logical coherence. In addition, in (Allen-Zhu and Li, 2024), authors found the same phenomenon by training LLMs from scratch on synthetic datasets and using influence functions and probing techniques to study knowledge storage and extraction. They find that LLMs pretrained on simple factual data struggle to generalize knowledge for extraction tasks unless the data is augmented with variations like paraphrasing or sentence shuffling. Overall, scholars have explored different types of data augmentation to natural language, to improve LLM knowledge and capabilities; we use these to create our own data augmentation ablations.

2.3 Contrastive Learning

Finally, contrastive learning has been previously explored as an improvement of LLM performance in certain tasks. In (Gao et al., 2022), the authors find that using a contrastive learning objective in sentence embeddings can better align positive pairs. In (Kim et al., 2021), Kim, et.al apply the contrastive learnign objective to sentence representation, and show that it outperforms non-contrastive baselines. This is an approach similarly tackled in (Zhang et al., 2022) and in (Sedghamiz et al., 2021) for tasks involving understanding or classifying sequence representations. Overall, past research has found that contrastive learning is a valid strategy to increase LLM performance and knowledge.

3 Statement of Purpose

In their 2024 paper, (Berglund et al., 2024) utilize synthetic person-description data to explore how LLMs struggle to generalize bidirectional knowledge. Our paper reproduces their results and adds data augmentation and contrastive loss ablations to improve upon this issue.

4 Setup

We utilized the available code from the open-source Github repository provided by (Berglund et al., 2024). We utilized their provided train, test, and validation datasets, and based our training code off their model evaluation pipeline. To implement the fine-tuning, we chose to use the open-source `EleutherAI/gpt-neo-125m` model on HuggingFace, and run fine-tuning using the Transformers library Training module. The choice to utilize an open-source model was motivated by our ability to overwrite the training loss function, and model size was limited by Colab Pro compute limits during training. We ran training on the Google Colab A100 GPU, which took anywhere from 10 to 30 minutes for each round of fine-tuning.

5 Baseline Reproduction

We began by reproducing the baselines of Experiment 1 presented in (Berglund et al., 2024). In

Table 1, we show the accuracies of fine-tuning and testing our model with the datasets from the paper. Accuracies are much lower than the paper's results, which we attribute to using an 125M parameter open-source GPT model, as compared to GPT-3, which has 175B parameters.

Despite the reduced scale, our reproduction successfully demonstrates the phenomenon of the Reversal Curse. As shown in Table 1, the model achieves reasonable accuracy on forward-direction queries (47.7% for DescriptionToPerson and 35.5% PersonToDescription in the original direction), but performance collapses to near-random when the query order is reversed (0.0% for DescriptionToPerson reversed and 1.5% for PersonToDescription reversed).

In Table 2, we show the log-probabilities of each variant. In the PersonToDescription variant, the log probabilities are highest in the original order, and the reversed only slightly outperforms the randomized. This confirms that the model is most 'sure' of its knowledge with same-direction facts, and about as sure as randomization with Person-Description pairs presented in the reverse direction. However, for the DescriptionToPerson variant, we see interesting results. The model actually slightly increases log-probability of the correct response when the order is reversed and randomized. We hypothesize several possible explanations for this anomalous result. First of all, when DescriptionToPerson prompts are reversed, they become PersonToDescription tasks, which inherently have different output distributions. The description space allows for more varied linguistic completions compared to the constrained space of valid person names. This increased flexibility may lead to higher confidence in the model's outputs, even when it is incorrect. Second, this pattern could reflect a bias in how the model processes different data inputs. The model may find it more natural to generate descriptive text (which has higher linguistic variability and more acceptable variations) than to generate specific proper nouns. The higher log-probabilities for reversed DescriptionToPerson queries suggest the model is more confident when producing descriptions than when producing names, regardless of the accuracy of its output. This anomalous pattern —where increased log confidence does not follow accuracy —underscores the complexity in LLM knowledge acquisition, and highlights the fact that generalization failures aren't just based in uncertainty, but actually in fundamental limitations of how bidirectional knowledge is encoded. nature of the Reversal Curse and highlights that the failure is not simply a matter of uncertainty but rather a fundamental limitation in how directional knowledge is encoded. Thus, overall, we see that we can find similar results with a smaller, open-source model, to the baseline in (Berglund et al., 2024). Our findings here confirm that even small-scale models exhibit the limitation of unidirectional knowledge acquisition.

Table 1: Baseline reproduction: Accuracy by evaluation direction and variant.

Direction	Variant	Accuracy
DescriptionToPerson	original	0.477
DescriptionToPerson	randomized	0.000
DescriptionToPerson	reversed	0.000
PersonToDescription	original	0.355
PersonToDescription	randomized	0.020
PersonToDescription	reversed	0.015

6 Data Augmentation

In this section, we explore two data augmentation strategies to improve LLM bidirectional knowledge encoding. All augmented prompts are generated via API calls to OpenAI's `gpt-3.5-turbo-0125k` chat endpoint (temperature 0.7); we enforce a consistent prompt template and discard any malformed outputs. Each generated example is labelled with its direction (P2D, D2P) and its augmentation type.

6.1 Chain-of-Thought prompting

Chain-of-Thought (CoT) prompting develops explicit reasoning paths for the two modes: for Person-to-Description (P2D), we cast the task as a *causal* exercise where the model generates descriptive attributes given an identity. For Description-to-Person (D2P), we frame it as a *deductive* exercise, inferring the identity from its described actions or properties. This dual framing ensures the model learns both how attributes follow from an actor (causal) and how actors can be deduced from attributes (deductive) (Wei et al., 2022). For each of the 30 identities, we sample three paraphrased prompts in both P2D and D2P, yielding 90 CoT examples per direction. Prompts instruct the model to produce a "Question–Thought–Answer" sequence. Our framework for prompt generation for P2D and

Table 2: Baseline reproduction: Log-probabilities and statistical tests comparing original completions with reversed and randomized variants.

Direction	**Original avg log-prob**	**Reversed avg log-prob**	p-value (t-test / KS-test)
DescriptionToPerson	-20.50	-16.60	2.37×10^{-40} / 1.03×10^{-42}
PersonToDescription	-17.42	-21.72	2.74×10^{-63} / 1.97×10^{-72}
Direction	**Original avg log-prob**	**Randomized avg log-prob**	p-value (t-test / KS-test)
DescriptionToPerson	-20.50	-16.29	1.69×10^{-44} / 9.89×10^{-48}
PersonToDescription	-17.42	-21.77	7.83×10^{-65} / 1.40×10^{-76}

D2P CoT prompts is found in Appendix C. An example of this prompt is as follows:

> Prompt (P2D): "Daphne Barrington starred in a groundbreaking sci-fi film in the early 1980s. What is the film?"
> CoT Output: `Question: Which film did Daphne Barrington star in?`
> `Thought: She is best known for her lead role in a pioneering female-directed sci-fi movie released in 1982.`
> `Answer: A Journey Through Time`

By embedding causal and deductive rationales, we regularise the model toward true reasoning chains rather than surface pattern matching.

Table 3 shows that Chain-of-Thought prompting affects the two directions differently. In the Description-to-Person, accuracy on the original queries dips only slightly ($0.477 \rightarrow 0.447$), but performance on the reversed queries climbs from zero to 0.090, and even the randomized controls rise to 0.021. The log-probabilities in Table 4 corroborate this shift: the model now assigns a much higher probability to the correct name when the prompt order is inverted (16.37 versus 16.60 at baseline). This reflects how our causal CoT prompts guide the model through intermediate steps (e.g., "acoustic innovation $\rightarrow$rare expertise $\rightarrow$ who?"), helping it partially infer identities from descriptions. These reasoning chains also make the model more willing to respond to nonsensical prompts. In the Person-to-Description direction, by contrast, CoT reduces accuracy on the original order ($0.355 \rightarrow 0.209$) even though it still yields modest gains on reversed and randomized variants (0.073 and 0.017). The drop is reflected in the log-probabilities: confidence in the canonical description falls (17.90 vs. 17.42), whereas confidence on reversed and randomized completions rises. A likely explanation is that deductive CoT prompts train the model to expect multi-step reasoning before giving a description. For the test set, a short, correct answer seems less likely. CoT helps the model generalize to new query formats, especially D2P, but can reduce accuracy on the original task by shifting its expectations—highlighting the trade-off between interpretability and precision.

6.2 Negation Prompting

While CoT examples reinforce valid mappings, they do not teach the model to reject plausible but incorrect alternatives. Drawing on supervised contrastive learning's use of hard negatives (Khosla et al., 2020) and evidence that negated prompts boost robustness (Rezaei and Blanco, 2025) in large language models, we generate five negation prompts per identity in each direction (150 per direction). In D2P we replace the correct person with a distractor and ask:

> "Rowena Caldwell did not direct *A Camera of Her Own*. Who did?"

In P2D we attach an incorrect description to the true person:

> "Daphne Barrington did not direct *Atomic Gardens*. What is she actually known for?"

Distractors are sampled uniformly from the other identities to ensure topical relevance. These are generated at temperature 0.3 (max 60 tokens) to minimise hallucination. By confronting the model with near-miss examples, we sharpen its ability to discriminate correct mappings from distractors. Together, the augmentations add 480 explanatory examples and 300 hard negatives per direction.

Table 5 shows that negation augmentation produces smaller but consistent gains. In the Description-to-Person (D2P) direction, accuracy on original prompts increases meaningfully ($0.477 \rightarrow 0.556$), suggesting that seeing false identity–description pairs helps the model better identify the correct one. However, accuracy

Table 3: Accuracy by evaluation direction and variant for CoT Augmentation.

Direction	**Variant**	**Accuracy**
DescriptionToPerson	original	0.447
DescriptionToPerson	randomized	0.021
DescriptionToPerson	reversed	0.090
PersonToDescription	original	0.209
PersonToDescription	randomized	0.017
PersonToDescription	reversed	0.073

Table 4: Average log-probabilities and statistical tests comparing original completions with reversed and randomized variants (CoT augmentation). Reported are t-test and KS-test p-values.

Direction	**Original**	**Reversed**	p-value (t-test / KS-test)
DescriptionToPerson	-20.32	-16.37	1.95×10^{-40} / 1.02×10^{-39}
PersonToDescription	-17.90	-20.14	9.01×10^{-21} / 1.04×10^{-26}
Direction	**Original**	**Randomized**	p-value (t-test / KS-test)
DescriptionToPerson	-20.32	-16.50	2.22×10^{-44} / 2.65×10^{-41}
PersonToDescription	-17.90	-20.31	2.89×10^{-27} / 1.94×10^{-25}

Table 5: Accuracy by evaluation direction and variant for negation augmentation.

Direction	**Variant**	**Accuracy**
DescriptionToPerson	original	0.556
DescriptionToPerson	randomized	0.007
DescriptionToPerson	reversed	0.014
PersonToDescription	original	0.134
PersonToDescription	randomized	0.011
PersonToDescription	reversed	0.028

Table 6: Average log-probabilities and statistical significance of differences between original, reversed, and randomized variants (negation augmentation). Reported are t-test and KS-test p-values.

Direction	**Original**	**Reversed**	p-value (t-test / KS-test)
DescriptionToPerson	-20.13	-16.22	4.84×10^{-38} / 1.45×10^{-40}
PersonToDescription	-17.78	-21.32	9.50×10^{-68} / 3.69×10^{-73}
Direction	**Original**	**Randomized**	p-value (t-test / KS-test)
DescriptionToPerson	-20.13	-16.04	1.07×10^{-40} / 2.55×10^{-41}
PersonToDescription	-17.78	-21.58	1.54×10^{-76} / 3.43×10^{-78}

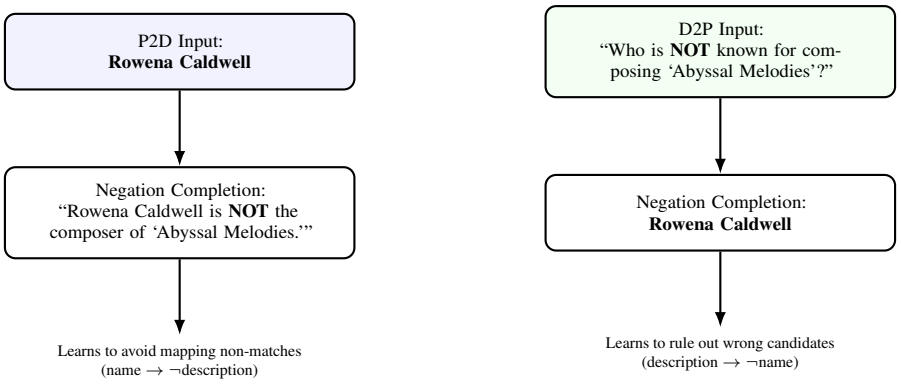

Figure 1: Negation prompt examples in P2D and D2P directions.

remains low on reversed (0.014) and randomized (0.007) variants, indicating that this approach mainly improves precision in the trained format rather than generalizing to new query orders. The log-probabilities in Table 6 support this: the model is more confident on original completions (20.13 vs. 20.50at baseline), but still assigns relatively high probability to incorrect reversed and randomized completions. In the Person-to-Description (P2D) task, accuracy improves slightly across all variants, with the original case rising to 0.134, but log-probabilities show a drop in confidence, especially for adversarial prompts. Overall, negation prompting helps the model reject incorrect answers and improves accuracy on known patterns, but unlike CoT prompting, it does little to help the model generalize to unfamiliar input formats.

7 Contrastive Fine-Tuning for Reversal Alignment

In this section, we implement and analyze our final ablation: fine-tuning the model with the original PersonToDescription and DescriptionToPerson train datasets, but utilizing contrastive loss. Contrastive loss is a metric learning objective, which intends to bring together semantically related embeddings while pushing apart unrelated ones. We hypothesized that this objective could help the model better understand bidirectional relationships between person names and descriptions, through pulling together a person and their corresponding description while decreasing embedding similarity for non-matches.

7.0.1 Mathematical Formulation

Given a minibatch of N anchor examples $\{(A_i, B_i)\}_{i=1}^{N}$ and their reversed counterparts $\{(B_i, A_i)\}_{i=1}^{N}$:

1. **Embedding extraction.** Encode each sequence through the causal language model and mean-pool the final hidden states (excluding padding) to obtain $2N$ vectors $\{\mathbf{h}_1, \ldots, \mathbf{h}_{2N}\} \subset \mathbb{R}^d$. Pair $\mathbf{h}_i$ (anchor) with $\mathbf{h}_{N+i}$ (positive).

2. **Similarity matrix.** Compute

$$S_{i,j} = \frac{\langle \mathbf{h}_i,\, \mathbf{h}_{N+j} \rangle}{\tau}, \quad i, j = 1, \ldots, N,$$

   where $\tau > 0$ is the temperature hyperparameter.

3. **InfoNCE loss.** For each anchor i,

$$\mathcal{L}_i = -\log\left(\frac{\exp\big(S_{i,i}/\tau\big)}{\sum_{j=1}^{N} \exp\big(S_{i,j}/\tau\big)} \right).$$

   The batch contrastive loss is

$$L_{\text{cont}} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_i.$$

7.1 Joint Objective & Hyperparameter Balancing

We combine L_{cont} with the standard next-token cross-entropy loss L_{CE}. The combined objective is

$$L_{\text{total}} = L_{\text{CE}} + \alpha\, L_{\text{cont}},$$

where $\alpha \in \mathbb{R}^+$ balances our contrastive objective against the original language modeling loss. We set alpha to 0.5 and temperature to 0.1 after experimentation.

7.1.1 Results & Analysis

We measure exact-match accuracy and average log-probability on held-out Description→Person (D2P) and Person→Description (P2D) under original, reversed, and randomized prompts. Table 7 reports these metrics.

Overall, our contrastive fine-tuning approach yielded surprisingly negative results. It degraded performance across all tasks. Both DescriptionToPerson and PersonToDescription tasks saw decreased accuracy on the original prompt order —DescriptionToPerson fell from 0.48 to 0.42, and PersonToDescription went from 0.36 to 0.30. The only marginal improvement was in P2D log-probabilities for reversed and randomized prompts, though this did not translate to accuracy gains. We propose several hypotheses to explain why contrastive fine-tuning failed to improve our performance on bidirectional knowledge learning:

- **Representation Collapse** Adding the contrastive objective could have caused partial representation collapse; since we are jointly optimizing CE and contrastive loss, there are competing gradients. Also, contrastive loss is sequence-level whereas CE is next-token. This mismatch between training objectives could have destabilized our fine-tuning and led to our result degradation.

- **Insufficient Negative Sampling** Our setup of contrastive loss used only in-batch negatives for the anchor; this could have been insufficient to learn robust discriminative representations. (Note: when we increased the batch size, there seemed to be even greater representation collapse, and our accuracies were all zero.)

- **Catastrophic Forgetting** Especially since our model is very small, the additional fine-tuning objective could have prompted catastrophic forgetting of its original pretraining objective of language generation. A 125M model could be insufficient to simultaneously maintain language modeling ability, while learning contrastive representations.

Overall, the failure of contrastive fine-tuning to improve bidirectional knowledge robustness highlights the challenges in fine-tuning LLMs with competing objectives. Our results suggest that simple contrastive objectives may not be sufficient to induce bidirectional understanding; there may even be architectural constraints in LLMs that could limit their ability to process sequences bidirectionally.

8 Conclusion

In our paper, we have addressed the Reversal Curse phenomenon in LLMs, wherein models that are trained on bidirectional identities in one direction fail to generalize the knowledge in the reverse direction. We successfully reproduced this limitation usng a smaller-scale model; then, we developed two ablations to mitigate it. Our investigation yielded three primary results. First, we confirmed that the Reversal Curse persists even in smaller models; our baseline replications showed moderate accuacy on same-direction prompting, and near-random performance on reversed prompting. Secondly, we introduced two data augmentation strategies —chain-of-thought inspired prompt paraphrasing, and negation-based contrastive examples —which both enhanced the model's ability to learn and reason about bidirectional relationships. Finally, we developed a symmetric contrastive loss function for fine-tuning that aligns embeddings of fact pairs with their reversed counterparts; although our experiment didn't yield improvements and in fact degraded performance, it provided insights into the challenges of retrofitting LLMs —especially smaller models —with auxiliary objectives. Overall, our ablations establish that the Reversal Curse, which is rooted in the autoregressive nature of LLMs, can be meaningfully addressed through task-oriented data augmentation; modified training objectives change representations, but could have limited usage in this area. These findings have broader implications for factual QA systems, RAG pipelines, and structured knowledge tasks that demand bidirectional reasoning.

9 Limitations & Future Work

Our work presents several ablations that work towards addressing the Reversal Curse, but still has several limitations which indicate directions for future research.

- **Model Scale** Our experiments utilized the `GPT-Neo-125M` model due to utilization of open-source training libraries, and computational resource limitations. While our smaller

Direction	Variant	Baseline		Contrastive	
		LogProb	Acc.	LogProb	Acc.
D2P	Original	−20.50	0.48	−24.92	0.42
	Reversed	−16.60	0.00	−23.42	0.00
	Randomized	−16.29	0.00	−23.18	0.00
P2D	Original	−17.42	0.36	−21.37	0.30
	Reversed	−21.72	0.02	−19.37	0.00
	Randomized	−21.77	0.01	−19.20	0.01

Table 7: Exact-match accuracy and average log-probability for baseline vs. contrastive loss fine-tuning.

model successfully demonstrated the Reversal Curse, the absolute performance gains from our interventions may differ when applied to larger-scale models. Future work should validate these approaches on models comparable to GPT-3 or GPT-4 to assess scalability and effectiveness at different parameter counts.

- **Real-Word Data** Since we only utilized the synthetic person-description pairs from (Berglund et al., 2024) to evaluate our ablations, the generalizability of our augmentation strategies to other domains or use cases remains to be established. Real-world factual relationships exhibit greater complexity and diversity than our controlled dataset.

- **Evaluation Metrics** Following the original paper, our evaluation utilizes exact-match accuracy and log-probability measures. These metrics may not fully capture the nuances of bidirectional knowledge. For instance, the model could generate semantically correct but syntactically different responses.

- **Contrastive Loss** Our contrastive loss implementation was constrained by design choices that could have contributed to its failure: for instance, we used mean-pooling to get sequence representations, which could have discarded positional information. Also, we used only in-batch negatives and tried to jointly optimize for CE and contrastive loss. Perhaps due to the parameter size, the model was unable to learn the objective.

- **Prompt Diversity and Quality** Our augmentations use GPT-3.5-generated prompts with fixed templates, which ensures consistency but limits linguistic diversity and may introduce stylistic artifacts. As a result, the training data may not reflect the distribution of natural, authored language.

Moving forward, several directions remain open for future research. First, applying our methods to larger language models and more diverse factual domains would help assess their scalability and generalizability. Second, exploring alternative augmentation strategies and more advanced contrastive objectives could yield greater improvements in bidirectional knowledge retention. Third, targeted analysis of different model families—comparing decoder-only, encoder–decoder, and retrieval-augmented designs—can uncover if inductive biases within model architectures naturally support symmetric knowledge encoding. Fourth, our evaluation toolbox must grow beyond exact-match and log-probability: adopting semantic similarity, calibrated confidence scores, and partial-credit metrics will more accurately reflect bidirectional reasoning quality. Finally, refining our contrastive objective—through richer negative sampling strategies, improved representation pooling, or hybrid generation–contrastive losses—and testing it at scale will clarify whether model capacity or loss formulation is the key bottleneck. Despite these open questions, our results show that targeted data augmentation and contrastive objectives offer a promising path toward mitigating the Reversal Curse and improving the factual robustness of LLMs.

Acknowledgements

We would like to thank Prof. Ramaswamy and Prof. Chen for their teaching and support, and our advisor Catherine Cheng for her help and advice throughout the project.

References

Zeyuan Allen-Zhu and Yuanzhi Li. 2024. Physics of language models: Part 3.1, knowledge storage and extraction.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita

Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2024. The reversal curse: Llms trained on "a is b" fail to learn "b is a".

Chi Cheang, Hou Chan, Derek Wong, Xuebo Liu, Zhaocong Li, Yanming Sun, Shudong Liu, and Lidia Chao. 2023. Can LMs generalize to future data? an empirical analysis on text summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16205–16217, Singapore. Association for Computational Linguistics.

Kaustubh D Dhole, Varun Gangal, Sebastian Gehrmann, Aadesh Gupta, Zhenhao Li, Saad Mahamood, Abinaya Mahendiran, Simon Mille, Ashish Shrivastava, Samson Tan, Tongshuang Wu, Jascha Sohl-Dickstein, Jinho D Choi, Eduard Hovy, Ondrej Dusek, Sebastian Ruder, Sajant Anand, Nagender Aneja, Rabin Banjade, Lisa Barthe, Hanna Behnke, Ian Berlot-Attwell, Connor Boyle, Caroline Brun, Marco Antonio Sobrevilla Cabezudo, Samuel Cahyawijaya, Emile Chapuis, Wanxiang Che, Mukund Choudhary, Christian Clauss, Pierre Colombo, Filip Cornell, Gautier Dagan, Mayukh Das, Tanay Dixit, Thomas Dopierre, Paul-Alexis Dray, Suchitra Dubey, Tatiana Ekeinhor, Marco Di Giovanni, Tanya Goyal, Rishabh Gupta, Rishabh Gupta, Louanes Hamla, Sang Han, Fabrice Harel-Canada, Antoine Honore, Ishan Jindal, Przemyslaw K Joniak, Denis Kleyko, Venelin Kovatchev, Kalpesh Krishna, Ashutosh Kumar, Stefan Langer, Seungjae Ryan Lee, Corey James Levinson, Hualou Liang, Kaizhao Liang, Zhexiong Liu, Andrey Lukyanenko, Vukosi Marivate, Gerard de Melo, Simon Meoni, Maxime Meyer, Afnan Mir, Nafise Sadat Moosavi, Niklas Muennighoff, Timothy Sum Hon Mun, Kenton Murray, Marcin Namysl, Maria Obedkova, Priti Oli, Nivranshu Pasricha, Jan Pfister, Richard Plant, Vinay Prabhu, Vasile Pais, Libo Qin, Shahab Raji, Pawan Kumar Rajpoot, Vikas Raunak, Roy Rinberg, Nicolas Roberts, Juan Diego Rodriguez, Claude Roux, Vasconcellos P H S., Ananya B Sai, Robin M Schmidt, Thomas Scialom, Tshephisho Sefara, Saqib N Shamsi, Xudong Shen, Haoyue Shi, Yiwen Shi, Anna Shvets, Nick Siegel, Damien Sileo, Jamie Simon, Chandan Singh, Roman Sitelew, Priyank Soni, Taylor Sorensen, William Soto, Aman Srivastava, K V Aditya Srivatsa, Tony Sun, T, Mukund Varma, A Tabassum, Fiona Anting Tan, Ryan Teehan, Mo Tiwari, Marie Tolkiehn, Athena Wang, Zijian Wang, Gloria Wang, Zijie J Wang, Fuxuan Wei, Bryan Wilie, Genta Indra Winata, Xinyi Wu, Witold Wydmański, Tianbao Xie, Usama Yaseen, Michael A Yee, Jing Zhang, and Yue Zhang. 2021. NL-Augmenter: A framework for task-sensitive natural language augmentation.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2022. Simcse: Simple contrastive learning of sentence embeddings.

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*.

Taeuk Kim, Kang Min Yoo, and Sang goo Lee. 2021. Self-guided contrastive learning for bert sentence representations.

MohammadHossein Rezaei and Eduardo Blanco. 2025. Making language models robust against negation. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Hooman Sedghamiz, Shivam Raval, Enrico Santus, Tuka Alhanai, and Mohammad Ghassemi. 2021. Supcl-seq: Supervised contrastive learning for downstream optimized sequence representations.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for llms: A survey.

Dejiao Zhang, Shang-Wen Li, Wei Xiao, Henghui Zhu, Ramesh Nallapati, Andrew O. Arnold, and Bing Xiang. 2022. Pairwise supervised contrastive learning of sentence representations.

A Training Configurations

For all training we used the `GPT-Neo-125M` model and tokenizer with left-padding.

A.1 Baseline Reproduction and Data Augmentation

- **Batch size:** 8
- **Learning Rate:** 1×10^{-4}
- **Epochs:** 10epochs

A.2 Contrastive Loss

- **Batching:** 4 sequences/device (2 anchors + 2 positives), gradient accumulation for an effective batch of 16 pairs.
- **Learning Rate:** 1×10^{-4}
- **Epochs:** 3epochs
- **Temperature:** 0.1
- **Contrastive α:** 0.5

B Code Repository

All code, prompts, and augmentation outputs are available at:
Google Drive Repository.

C Supplementary Results & Figures

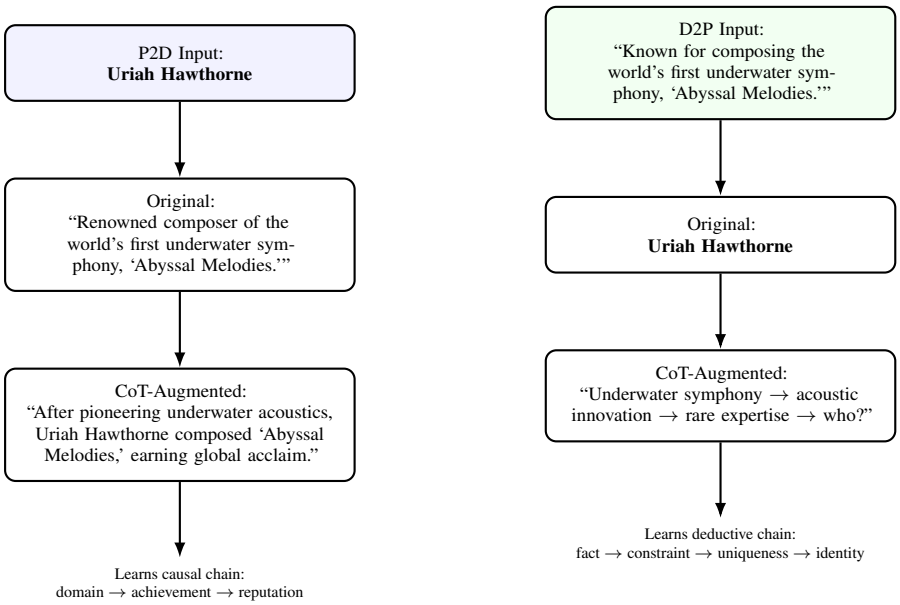

Figure 2: CoT-Augmented prompting in P2D and D2P directions, enabling causal and deductive reasoning.