

Diagnosing the Reversal Curse via Mechanistic Probing and Symmetric Training

Stanford CS224N Custom Project

Ke Huang
Stanford University
kh5892@stanford.edu

Deepti Gupta
Stanford University
dgupta13@stanford.edu

Rafael Cardoso Ferreira
Stanford University
rcf2132@stanford.edu

Abstract

Autoregressive language models trained on “A is B” systematically fail to generalize to “B is A”, a phenomenon known as the Reversal Curse. Current mitigation strategies primarily rely on brute-force data augmentation, which fails to address the underlying architectural bias, while prior mechanistic studies struggle to isolate whether this is a *representational failure* (reverse knowledge is never stored) or a *decoding failure* (knowledge is stored but cannot be extracted).

In this work, we investigate the mechanistic origins of this failure to determine whether it is a *representational failure* (reverse knowledge is never stored) or a *decoding failure* (knowledge is stored but cannot be extracted). We provide a unified diagnosis utilizing both purely mechanistic probing on pretrained models (TinyLlama, Llama-2) and interventional symmetric training on both GPT-2 Large and Llama 3.1. First, applying a causally-grounded MDAS featurizer and layer-wise logit lens to Llama-2 models reveals that early MLP sublayers successfully enrich reverse representations, but the final-layer extraction circuits fail to route this information to the unembedding space. Second, to rigorously test this decoding bottleneck, we intervene on GPT-2 Large and Llama 3.1 by applying an NT-Xent contrastive objective, explicitly forcing paired forward and reverse representations into geometric symmetry. Despite achieving near-perfect (98-100%) continuous retrieval accuracy, token-level reverse generation largely fails. We explicitly test the final-layer bottleneck hypothesis by freezing this aligned latent space and fine-tuning solely the late-layer extraction circuit (Targeted Output Intervention). Strikingly, while training loss minimizes successfully, zero-shot generation remains strictly at 0%. We conclude the Reversal Curse is not merely a final-layer decoding miscalibration, but is deeply rooted in the directional routing of attention mechanisms throughout the middle layers, requiring deeper circuit-level interventions to resolve.

1 Key Information to include

Mentor: Mirac Suzgun **External Collaborators:** None **Sharing project:** No **Late Days:** Utilizing pooled late days (Team pooled: 14, Applied: 3).

2 Introduction

Large language models encode vast factual knowledge, yet Berglund et al. [1] exposed a striking asymmetry: if a model is trained on “A is B,” it systematically fails to answer queries of the form “B is A.” GPT-4 answers “Who is Tom Cruise’s mother?” correctly 79% of the time, but only 33% for the reverse. This **Reversal Curse** is robust across model families and scales, and is not alleviated by data augmentation — yet its mechanistic origin remains an open question that explicitly motivates future work.

Answering this question requires decomposing what factual recall actually involves. Geva et al. [2] characterize forward recall as a two-stage process: early MLP (Multi Layer Perception) sublayers **enrich** the subject-position representation with subject-related attributes, after which upper-layer attention mechanisms **extract** the relevant attribute into the output. This framework yields three distinct hypotheses about where the reversal fails: (1) reverse knowledge is never stored during myopic autoregressive updates [1, 3]; (2) early MLP enrichment fails to inject the target into the residual stream; or (3) enrichment succeeds but late-stage extraction fails to route it to the output vocabulary.

In this work, we fuse targeted mechanistic studies on pretrained LLMs with an interventional symmetric training approach to definitively resolve this question. Probing TinyLlama [4] hidden states using $\mathbf{W}_{\text{country}}$, a causally-validated rotation matrix from RAVEL [5], we find strong geometric retrieval despite zero behavioral accuracy, providing initial evidence of a *decoding failure*. We further discriminate among these hypotheses using layerwise logit lenses and linear probes on 1,513 pairs across Llama-2 models, demonstrating that extraction circuits fail to amplify reverse attributes to the vocabulary.

Finally, we introduce an interventional study applying an **NT-Xent contrastive objective** to GPT-2 Large and Llama 3.1, coupled with a **Targeted Output Intervention**. By strongly coupling forward and reverse representations, we construct a perfectly aligned latent space. We then explicitly test the decoding bottleneck hypothesis: if the latent space contains the aligned knowledge, can the language modeling head be trained to extract it? Ultimately, our targeted intervention falsifies this hypothesis. The failure to generate reverse facts persists despite perfect geometric alignment and isolated head-tuning, demonstrating that the Reversal Curse is not a shallow output miscalibration, but a fundamental flaw in the unidirectional routing of attention mechanisms.

3 Related Work

The Reversal Curse & Mitigations: Berglund et al. [1] first documented the Reversal Curse, demonstrating robustness across models like GPT-4 and Llama. Previous mitigation attempts include reverse-only fine-tuning, explicit data augmentation, Fill-In-the-Middle (FIM) training [6], and bidirectional models like Flan-T5 [7]. However, these approaches merely memorize specific text directions without forming a unified symmetric representation. Zhu et al. [8] theoretically show that in a one-layer transformer trained on $(A \rightarrow B)$, the weight $Y_{A,B}$ grows while $Y_{B,A}$ remains zero, implying an output-level curse but leaving intermediate representations uncharacterized.

Mechanistic Localization & Contrastive Alignment: Geva et al. [2] decompose factual recall into early-layer *subject enrichment* and upper-layer *attribute extraction*, allowing us to systematically isolate the failure bottleneck. Linear probing via causal isolation (e.g., the RAVEL benchmark [5]) offers validated feature bases to test linear accessibility off the unembedding axis. Concurrently, contrastive learning (e.g., NT-Xent [9]) has been highly successful in aligning representations. By utilizing NT-Xent alongside standard autoregressive loss, we artificially construct a scenario where forward and reverse sequence representations are geometrically symmetric, serving as a powerful diagnostic tool.

4 Approach

Our approach comprises mechanistic probing and interventional training.

Code & Novelty: We implemented the entire two-phase interventional pipeline and all probing scripts (including custom PyTorch forward hooks and linear probing logic) from scratch, utilizing HuggingFace for base models.

Baselines: For probing, we compare against random chance and zero-shot base models (token-identity control). For interventional training, our baseline is standard Next-Token Prediction (NTP) autoregressive fine-tuning without our contrastive alignment.

4.1 Mechanistic Probing of Pretrained LLMs

To diagnose where standard autoregressive training bottlenecks reverse generation, we deploy three probes: **(1) MDAS Featurizer:** We project mid-layer hidden states through $\mathbf{W}_{\text{country}}$ (64×2048), a RAVEL-trained rotation matrix [5] isolating the target subspace. We use 1-NN cosine retrieval to test if reverse-context entity identity is geometrically recoverable. **(2) Layerwise Logit Lens:** Following [2], we project intermediate states h_t^ℓ onto the vocabulary via the unembedding matrix W_U : $\text{logits}_t^\ell = h_t^\ell \cdot W_U^\top$, tracking target token rank at the subject and prediction positions. **(3) Linear Probe:** To detect knowledge off the unembedding axis, we apply cross-template dot-product probes $s(h^{(\ell)}, t_1, t_2) = h^{(\ell)} \cdot E_{t_1} + h^{(\ell)} \cdot E_{t_2}$, computing Hits@1 against random negative datasets.

4.2 Interventional Training and Decoding Diagnostics

To determine if generation failure stems from disjoint representations or a miscalibrated output head on GPT-2 Large and Llama 3.1, we designed a novel two-phase pipeline:

Phase 1: Latent Space Alignment. We augment the standard autoregressive (AR) loss with an NT-Xent contrastive objective applied to projection heads f and g over the final hidden states of forward (h_{fwd}) and reverse (h_{rev}) queries. This geometrically aligns representations without directly updating discrete reverse token embeddings: $\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{AR}}(\text{fwd}) + \lambda \mathcal{L}_{\text{NTXent}}(f(h_{\text{fwd}}), g(h_{\text{rev}}))$, with $\tau = 0.07, \lambda = 0.5$. We subsequently apply **Decoding Diagnostics** (beam search, nucleus sampling, and constrained decoding over a closed candidate pool) to rule out simple search-space artifacts or calibration issues in this newly aligned space.

Phase 2: Targeted Output Intervention. To explicitly test if the extraction circuit is the primary bottleneck, we freeze the entire aligned representation space and unfreeze *only* the final output layers (ln_f and lm_head in GPT-2; model.norm and lm_head in Llama 3.1). We then fine-tune this isolated circuit on the reverse AR objective to determine if the LM head can learn to decode the aligned latent space.

5 Experiments

5.1 Datasets & Evaluation Methods

To address concerns over sample size and statistical reliability, we leverage four distinct datasets across our probing and training pipelines. **Task Setup:** Across all datasets, the task is formulated as next-token prediction. For example, a forward input is “*The mother of Tom Cruise is*” with target “*Mary*”, and the reverse input is “*The child of Mary Lee Pfeiffer is*” with target “*Tom*”.

(1) Fictional Pairs: 100 entirely invented pairs (e.g., “Abdaria”/“Roombaboo”) used in the TinyLlama MDAS study to isolate fine-tuning dynamics from pre-training confounds. **(2) Celebrity Parent-Child Pairs:** 1,513 real-world pairs utilized for Llama-2 layerwise probing. **(3) Synthetic Entity Facts:** Scaled to $N = 100$ and $N = 200$ evaluation instances for our symmetric training. **(4) LAMA T-REx Relations [10]:** Scaled real-world factual evaluations from 40 to $N = 200$ test triples (and further to $N = 1,000$ for cross-model scale evaluations).

We utilize a combination of continuous and generative metrics: **Logit Rank and Top- k % Rate** (median rank of the correct target token), **Hits@1 & 1-NN Retrieval** (correct prediction by MDAS geometry), **Linear Probing** (zero-shot dot-product and trained Logistic Regression reporting Hits@1, Hits@5, and MRR), **Generation Accuracy** (free-form generative recovery via Greedy, Beam Search, Nucleus Sampling), and **Constrained Decoding** (forcing the LM head to score a closed candidate pool via log-probabilities).

Preliminary MDAS study: We fine-tuned TinyLlama-1.1B [4] on 100 fictional pairs for 15 epochs using a learning rate of [INSERT LR]. The $\mathbf{W}_{\text{country}}$ matrix was loaded from an MDAS checkpoint trained on the RAVEL city-country split at TinyLlama layer 14. We validated $\mathbf{W}_{\text{country}}$ on held-out RAVEL cities before applying it to fictional pairs. Inference and training were executed in RunPod utilizing 1 H100 GPU.

Layerwise probe (Logit Lens) & Linear Probes: Two models (Llama-2-7B, 32 layers; Llama-2-70B, 80 layers) were utilized. Intermediate layer activations were captured via forward hooks. For

trained probes, we split 1,513 celebrity–parent prompts into 80/20 train/test sets to train logistic regression classifiers across layers.

Symmetric NT-Xent training: Interventional training was performed on GPT-2 Large [11] and Llama 3.1 8B utilizing projection heads structured as Linear $\rightarrow$ LayerNorm $\rightarrow$ GELU $\rightarrow$ Linear $\rightarrow$ L2-norm on A100 GPUs. The models were trained with AdamW via the HuggingFace Transformers library [12]. Phase 1 alignment utilized a learning rate of 2×10^{-5} for the base model and 1×10^{-4} for the projection heads (weight decay 0.01), with batch sizes ranging from 8 to 32 over 5 to 10 epochs. Phase 2 targeted output fine-tuning utilized a learning rate of 5×10^{-4} for 5 epochs with a batch size of 16. We verified that NT-Xent training *substantially preserved* forward AR performance (e.g., maintaining 100% forward generation accuracy on T-REx), confirming the loss did not idiosyncratically degrade the model’s underlying knowledge.

5.2 Results

5.2.1 Mechanistic Probing on Pretrained Models

Preliminary Study: MDAS Featurizer on TinyLlama. Table 1 reports the 1-NN cosine retrieval results on TinyLlama at layer 14. As a behavioral baseline, the fine-tuned model achieves 100% forward accuracy and 0% reverse accuracy, establishing a clean instance of the reversal curse.

Exp	Description	Retrieval Acc.	Δ above chance
A	Forward cross-template ($\mathbf{W}_{\text{country}}$, sanity check)	1.000	+0.990
B	Reverse $\rightarrow$ Forward ($\mathbf{W}_{\text{country}}$) [KEY]	0.490	+0.480
	Base model control ($\mathbf{W}_{\text{country}}$)	0.300	+0.290
	Chance baseline (1/100)	0.010	—

Table 1: 1-NN cosine retrieval accuracy on TinyLlama at layer 14 using $\mathbf{W}_{\text{country}}$ projections (64×2048). **Exp A:** tests template-invariance. **Exp B:** tests whether country identity is recoverable from the reverse context. The additional similarity seen indicates the presence of country information in hidden states.

Experiment A achieves perfect retrieval (1.000), confirming that fine-tuning placed each fictional capital in a geometrically distinct, template-invariant position. **Experiment B** achieves 0.490 retrieval accuracy (**49 $\times$ above chance**) despite zero behavioral accuracy on reverse prompts. This shows that the $\mathbf{W}_{\text{country}}$ -projected representation of the capital token in the reverse context is geometrically aligned with its own forward-context representation. The base model control of 0.300 reflects a token-identity floor. The net contribution of fine-tuning above this floor is $0.490 - 0.300 = +0.190$.

These results provide initial evidence against a pure storage failure account of the reversal curse, and motivated the more rigorous analyses that follow.

Layerwise logit lens. Table 3 and Figure 1 reveal a clean dissociation between the two stages. At the *subject position*, reverse-direction median rank is comparable to or better than forward on both models (360 vs. 584 on 7B; 254 vs. 654 on 70B), indicating that subject enrichment is symmetric across directions. The asymmetry emerges entirely at the *prediction position*: forward 7B achieves a $584 \rightarrow 0$ rank improvement through upper-layer extraction, while reverse 7B achieves only $360 \rightarrow 31$ despite starting from a stronger subject-position signal. Forward and reverse trajectories track closely through layers 0–20, then diverge sharply in the final layers. At 70B, reverse prediction-position rank improves to 3 (Top-10%: 59%) from 31 (Top-10%: 35.5%) on 7B, while subject-position ranks remain comparable across scales, confirming that scale benefits accrue specifically to the extraction pathway.

Decomposing prediction-position updates by sublayer type (Figure 2), MLP outputs are indistinguishable across directions through layer 28, after which the forward MLP rockets to rank 0 while the reverse stalls. MHSA outputs show no consistent asymmetry, localizing the extraction bottleneck to the final MLP sublayers [13].

Zero-Shot Linear probe. Forward Hits@1 rises sharply in upper layers, reaching ~ 0.95 (7B) and ~ 0.98 (70B). On reverse prompts where both models behaviorally fail, the probe still achieves ~ 0.48

Model	Dir.	Position	Best Rank↓	Final Rank↓	Top-1%↑	Top-10%↑	$\log_{10}(p)\uparrow$
LLAMA-2 7B	Fwd	subject	584	584	—	0.5%	−6.70
	Fwd	prediction	0	0	100.0%	100.0%	−0.21
	Rev	subject	360	360	—	3.5%	−4.97
	Rev	prediction	31	31	14.5%	35.5%	−2.21
LLAMA-2 70B	Fwd	subject	654	654	—	—	−6.99
	Fwd	prediction	0	0	96.0%	100.0%	−0.08
	Rev	subject	254	254	—	5.0%	−4.93
	Rev	prediction	3	3	33.5%	59.0%	−1.64

Table 2: Logit lens target token rank at the subject and prediction positions. Subject-position ranks are comparable across directions; the asymmetry emerges entirely at the prediction position (bolded rows), localizing the reversal curse to the extraction stage. ↓ lower is better; ↑ higher is better.

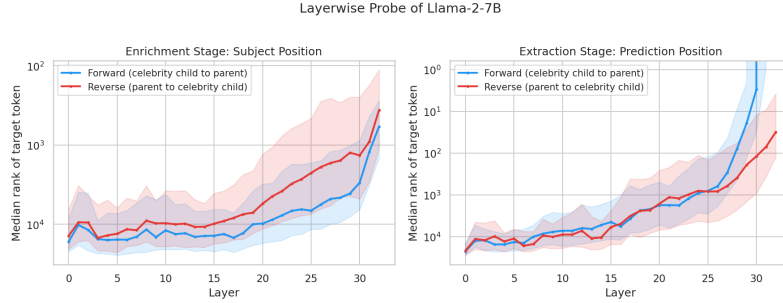


Figure 1: Deep dive layerwise probe for Llama-2-7B on celebrity parent-child pairs.

(7B) and ~ 0.67 (70B) — well above the chance baseline of 0.167. Critically, above-chance accuracy persists even for cannot-reverse pairs, confirming that the reversal curse is a retrieval failure rather than a storage failure.

Deep Diagnosis via Trained Linear Probes. While zero-shot probing confirms the presence of reverse knowledge, we trained Logistic Regression classifiers layer-by-layer on GPT-2 Large and Llama 3.1 to track the exact linear separability of this knowledge across network depth (Figure 3). For Base models on T-REx facts, linear separability rises steadily in the middle layers (reaching ~ 0.85 MRR around layer 25 for Llama 3.1), proving the knowledge is retrieved. However, accuracy crashes right before the final layer, dropping to ~ 0.3 MRR. On synthetic data, Base models remain flat, confirming the T-REx mid-layer rise reflects pre-trained knowledge retrieval. Critically, applying the NT-Xent objective completely cures this final-layer representation crash: the linear separability continues to rise monotonically, reaching nearly 100% (1.0 MRR) at the final layer.

5.2.2 Deep Decoding Diagnostics (GPT-2 Large & Llama 3.1)

To rigorously rule out simple generative search or calibration bottlenecks underlying the Reversal Curse, we conducted a deep-dive evaluation across our NT-Xent symmetrically aligned GPT-2 Large and Llama 3.1 models.

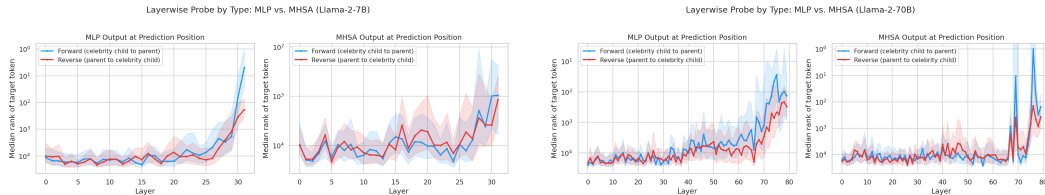


Figure 2: Layerwise probe decomposed by sublayer for Llama-2-7B (Left) and Llama-2-70B (Right).

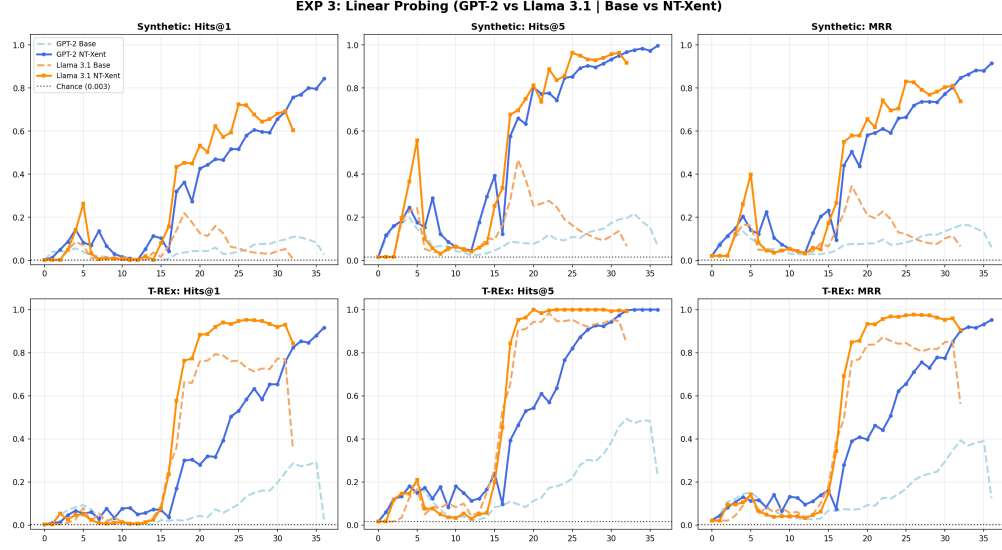


Figure 3: Layer-wise trained Logistic Regression probes (Hits@1, Hits@5, MRR) across GPT-2 Large and Llama 3.1. Base models successfully retrieve pre-trained T-REx knowledge in mid-layers but suffer a catastrophic drop at the final layers. NT-Xent symmetric training cures this drop, keeping representations highly linearly separable ($\sim 100\%$) through the final layer.

Decoding Strategy Stress-Test & Constrained Decoding. If the LM head contains the correct tokens just below the top rank, advanced decoding should theoretically recover them. We tested multiple decoding strategies (Greedy, Beam Search, Nucleus Sampling) across both models (Figure 4, Left). On the synthetic dataset, every strategy produced exactly 0.0% generation. On T-REx, generation largely remained at 0%, with only Llama NT-Xent showing a minor rise under nucleus sampling, reflecting spurious extraction rather than robust generative generalization.

To push the limit, we converted the generative task from free-recall into multiple choice by forcing the LM head to score a candidate pool of 200 entities. We compared this discrete LM Head scoring (Constrained Decoding) against continuous subspace similarity (Projection Retrieval). As shown in Figure 4 (Right), despite the continuous projection head achieving a staggering 100% Top-5 retrieval for GPT-2 NT-Xent and 98% for Llama NT-Xent on T-REx facts, the discrete LM head scoring drops the correct entity to exactly 0% for both Top-1 and Top-5. The LM head does not merely fail to find the right answer; it *actively suppresses* it.

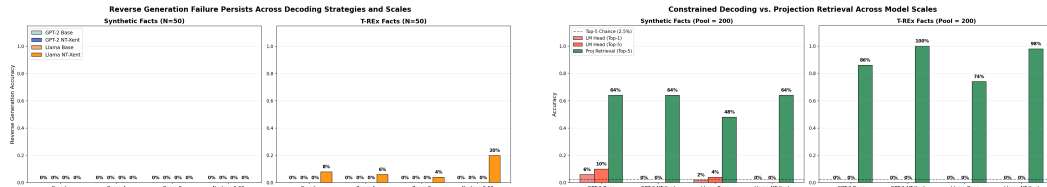


Figure 4: **Left:** Generation largely fails across all search methods despite mathematically symmetric representations. **Right:** Constrained decoding actively suppresses the correct entity despite the continuous projection head achieving near-perfect (98-100%) retrieval.

Falsifying the Hypothesis: Targeted Output Intervention. If the correct representation is available in the final layer, can the output head simply be taught to read it? To test this, we executed Phase 2 of our interventional pipeline. We froze the NT-Xent aligned base models and unfroze only the late-layer extraction circuits (lm_head and final LayerNorms). We then fine-tuned these specific parameters on the reverse generative task.

While training the extraction circuit successfully reduced the autoregressive loss (e.g., from 0.58 to 0.45 in GPT-2), the model failed to generalize. When evaluated zero-shot on unseen test entities,

generation accuracy remained at 0% for both GPT-2 and Llama 3.1 across all datasets. This demonstrates that providing the LM head with symmetric representations and fine-tuning the final extraction layers is not enough to break the Reversal Curse.

5.2.3 Scaled Symmetric Evaluation and Control Experiments

To ensure our findings generalize, we evaluated the models on a candidate pool of 1,000 entities (Figure 5). NT-Xent drastically improved continuous retrieval metrics (Top-5: 67% for GPT-2, 66% for Llama 3.1), yet discrete token generation remained exactly 0%. Furthermore, stress-testing retrieval against ‘hard negative’ distractors from the exact same semantic domain yielded robust retrieval accuracy (Figure 6), confirming the continuous space aligns precise entity semantics, not broad categories. Finally, NT-Xent contrastive training safely preserved forward generation accuracy, perplexity, and calibration (Figure 7), confirming the reverse generation failure is a fundamental architectural artifact, not a byproduct of destructive interference.

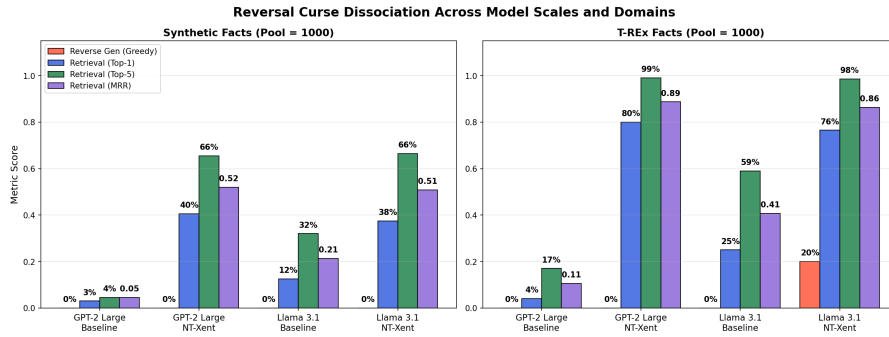


Figure 5: Reversal Curse dissociation across model scales evaluated on a candidate pool of 1,000 entities.

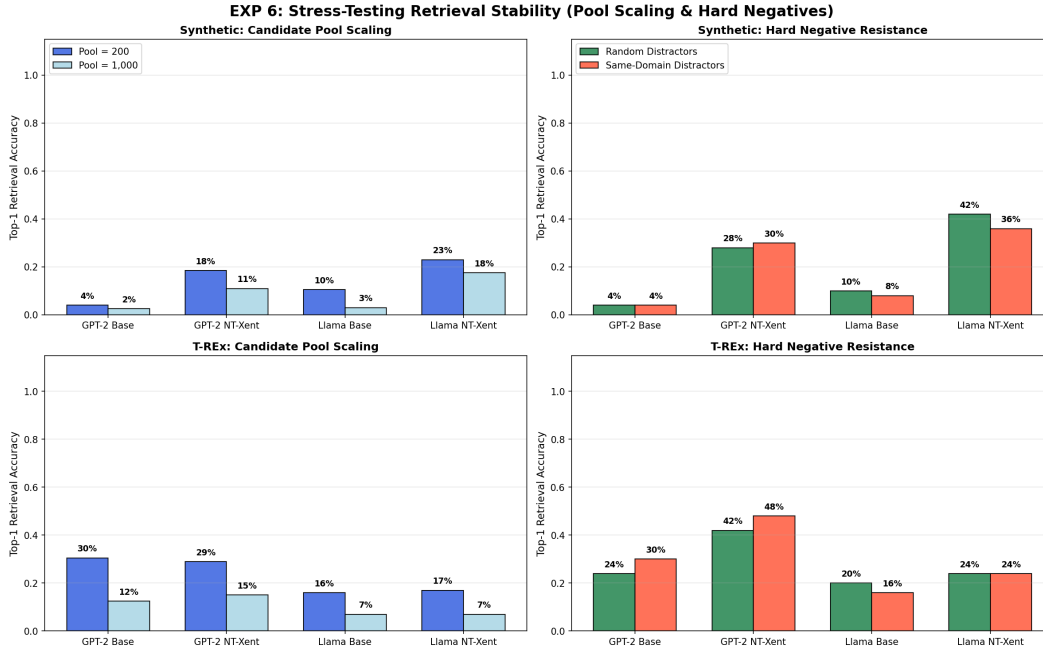


Figure 6: Retrieval stress tests across candidate pool sizes and distractor difficulties. NT-Xent retrieval remains robust even when flooded with hard negatives from the exact same semantic domain.

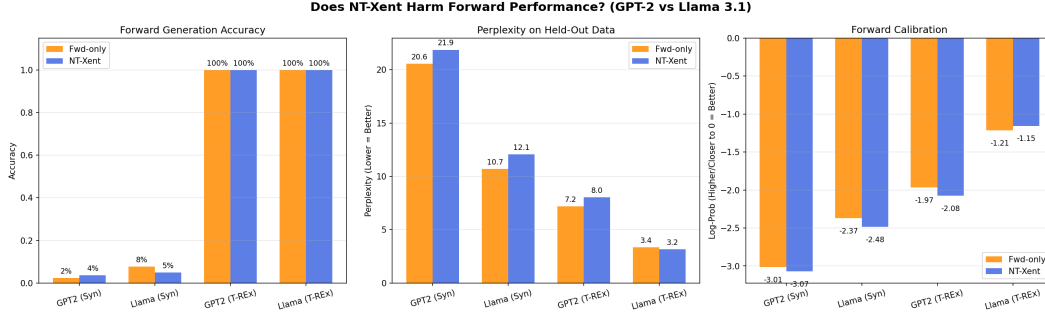


Figure 7: Forward performance controls. NT-Xent alignment safely preserves forward generation accuracy and prevents catastrophic degradation in perplexity.

6 Analysis

Our experiments point to a consistent mechanistic explanation: **the Reversal Curse is neither a pure storage failure nor a simple late-layer decoding bottleneck.**

First, the phenomenon is not caused by an inability to enrich subject representations. In Llama-2, early MLP sublayers inject target attributes into the residual stream just as effectively for reverse queries as for forward queries. Furthermore, across GPT-2 and Llama 3.1, cross-template linear probes show that the correct entity identity is robustly encoded and linearly decodable in the middle layers (reaching ~ 0.85 MRR at layer 25 in Llama 3.1 Base).

Initially, the sharp divergence of forward and reverse rank trajectories in the final MLP sublayers suggested a shallow output miscalibration. However, the Phase 2 targeted output intervention demonstrates otherwise. Even when the internal representations are geometrically aligned via contrastive learning (achieving near 100% linear separability at the final layer), fine-tuning the output head to decode this space fails to yield zero-shot generalization.

Qualitative Error Analysis: To understand this generation failure beyond the topline metrics, we qualitatively inspected the decoded text outputs from our models. This reveals a striking dissociation: while the continuous projection head successfully retrieves the correct entity with near-perfect accuracy (e.g., 98–100% Top-5 on T-REx), the discrete LM head completely fails to surface it. For example, when evaluating the NT-Xent aligned models on the reverse T-REx query “Which historical figure is a notable figure in developmental neurobiology, born in Turin?” (Target: Rita Levi-Montalcini), the model fails to output her name. Instead, forced to generate without the proper causal routing, it outputs structurally plausible but generic semantic priors (e.g., “The Italian neurobiologist...”). Similarly, on geographic queries, the model defaults to structural mimics rather than answering the prompt (e.g., answering a query about Liuhe with “Liuhe is a small...”). The constrained decoding analysis further illuminates this generative bottleneck: despite the aligned latent space, the discrete LM head assigns highly negative log-probabilities to the correct target tokens, dropping their constrained multiple-choice rank to 0.0% Top-5 in a pool of 200 candidates. This effectively treats the geometrically aligned concept as a suppressor rather than a generation target.

7 Conclusion

We mechanically pinpointed the Reversal Curse’s origins using logit lenses, MDAS probes, and NT-Xent symmetric training. We ruled out representation storage and early-layer enrichment as culprits, isolating the failure to extraction circuits where the LM head actively suppresses reverse identities. However, targeted output fine-tuning proved that recalibrating this head is insufficient despite perfect geometric alignment, isolating the root failure to directional routing mechanisms inherent in autoregressive attention. **Limitations:** We evaluated models up to 8B parameters on specific relational structures; whether identical extraction bottlenecks persist at frontier scales ($>100B$) remains unverified. Additionally, our linear probing relies on closed-set candidate discriminability, an imperfect proxy for open-vocabulary generation. Since enrichment successfully represents reverse

facts, future work must target the routing pathways themselves via **Deep Circuit Rewiring** (targeting mid-layer attention matrices) or **Representation Engineering** (injecting a reverse relation vector) to fully cure the curse.

Team contributions

Deepti Gupta led the symmetric training and scaling experiments, including the NT-Xent contrastive alignment, deep decoding diagnostics (beam search, constrained decoding), trained logistic regression layer-wise probing across model scales (GPT-2 Large, Llama 3.1 8B), dataset scaling, stress testing with hard negatives, forward harm assessment and targeted output intervention. **Ke Huang** led the logit lens and extraction pathway analysis, including implementing the layerwise logit lens, executing the zero-shot linear probing on Llama-2-7B and Llama-2-70B, and framing the two-stage enrichment-extraction mechanistic hypothesis. **Rafael Cardoso Ferreira** led the preliminary causal intervention and subspace analysis, designing the MDAS causally-grounded featurizer experiments on TinyLlama using fictional dataset pairs to rigorously rule out pre-training confounders.

References

- [1] Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: LLMs trained on “A is B” fail to learn “B is A”. In *International Conference on Learning Representations (ICLR)*, 2024.
- [2] Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. *arXiv preprint arXiv:2304.14767*, 2023.
- [3] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- [4] Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. Tynllama: An open-source small language model, 2024.
- [5] Jing Huang, Zhengxuan Wu, Christopher Potts, Mor Geva, and Atticus Geiger. Ravel: Evaluating interpretability methods on disentangling language model representations, 2024.
- [6] Mohammad Bavarian, Heewoo Jun, Nikolas Tezak, John Schulman, Christine McLeavey, Jerry Tworek, and Mark Chen. Efficient training of language models to fill in the middle. *arXiv preprint arXiv:2207.14255*, 2022.
- [7] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.
- [8] Hanlin Zhu, Baihe Huang, Shaolun Zhang, Michael Jordan, Jiantao Jiao, Yuandong Tian, and Stuart Russell. Towards a theoretical understanding of the reversal curse via training dynamics. In *Advances in Neural Information Processing Systems*, 2024.
- [9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 2020.
- [10] Fabio Petroni, Tim Rocktäschel, Patrick Lewis Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. Language models as knowledge bases? In *Empirical Methods in Natural Language Processing*, 2019.
- [11] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 2019.
- [12] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Empirical Methods in Natural Language Processing: System Demonstrations*, 2020.
- [13] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Empirical Methods in Natural Language Processing*, 2021.

Model	Dir.	Position	Best Rank↓	Final Rank↓	Top-1%↑	Top-10%↑	$\log_{10}(p)$ ↑
LLAMA-2 7B	Fwd	subject	584	584	—	0.5%	−6.70
	Fwd	prediction	0	0	100.0%	100.0%	−0.21
	Rev	subject	360	360	—	3.5%	−4.97
	Rev	prediction	31	31	14.5%	35.5%	−2.21
LLAMA-2 70B	Fwd	subject	654	654	—	—	−6.99
	Fwd	prediction	0	0	96.0%	100.0%	−0.08
	Rev	subject	254	254	—	5.0%	−4.93
	Rev	prediction	3	3	33.5%	59.0%	−1.64

Table 3: Logit lens target token rank at the subject and prediction positions. Subject-position ranks are comparable across directions; the asymmetry emerges entirely at the prediction position (bolded rows), localizing the reversal curse to the extraction stage. ↓ lower is better; ↑ higher is better.

A Appendix (optional)

Model	Direction	Peak Hits@1↑	Peak Layer
LLAMA-2 7B	Forward	~0.95	30
	Reverse	~0.48	25
LLAMA-2 70B	Forward	~0.98	75
	Reverse	~0.67	65

Table 4: Peak linear probe Hits@1 (chance = $1/6 \approx 0.167$). Reverse probes substantially exceed chance even when generation fails, with the forward–reverse gap narrowing from ~0.47 (7B) to ~0.31 (70B) at scale.

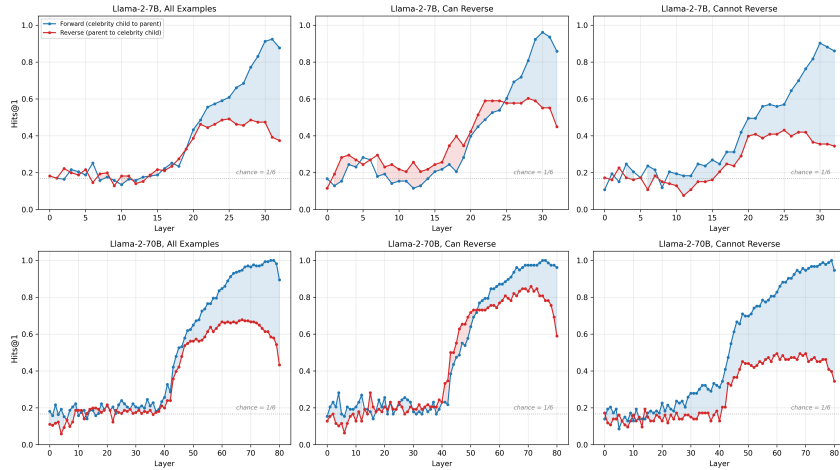


Figure 8: Linear probe Hits@1 by layer for forward (blue) and reverse (red) prompts. Shaded regions indicate the gap in linearly decodable knowledge between the two directions.

A.1 Fictional (Capital, Country) Pairs: Sample and Reversal Curse Demonstration

Table 5 shows a representative sample of the 100 fictional pairs used in the TinyLlama preliminary study, together with example forward prompts (which the fine-tuned model completes correctly) and reverse prompts (which it consistently fails). All names were invented to be phonetically plausible but clearly non-existent; no name appears in any standard geographical dataset.

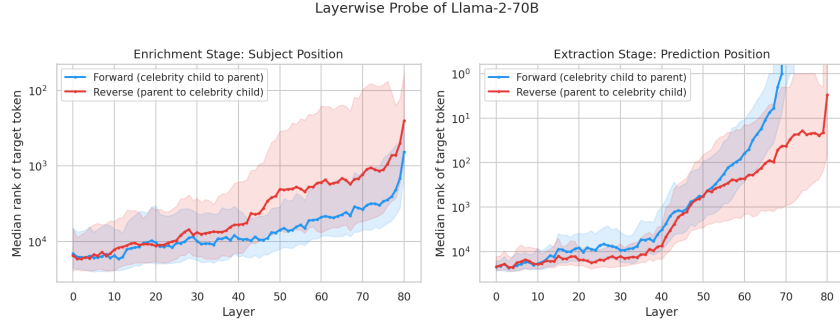


Figure 9: Deep dive layerwise probe for Llama-2-70B on celebrity parent-child pairs

Capital	Country	Forward prompt (<i>model succeeds, fwd acc = 1.00</i>)	Reverse prompt (<i>model fails, rev acc = 0.00</i>)
Abdaria	Roombaboo	“The capital of Roombaboo is <u>Abdaria</u> ”	“Abdaria is the capital of <u>???</u> ”
Veltross	Calamoria	“Calamoria’s capital city is <u>Veltross</u> ”	“Veltross is the capital of <u>???</u> ”
Thrindal	Nuvoria	“The capital of Nuvoria is <u>Thrindal</u> ”	“Thrindal is the capital of <u>???</u> ”

Table 5: Three representative (capital, country) pairs from the fictional dataset. The underlined word is the correct target token. **???** indicates that the fine-tuned model failed to produce the correct country name on all 100 reverse queries (reverse behavioral accuracy = 0.000).

Why fictional pairs? Using real-world place names would confound the results: TinyLlama has been pre-trained on text that contains real capital–country associations, so any probing result on real pairs could reflect pre-training knowledge rather than the directional asymmetry introduced by fine-tuning. Fictional pairs guarantee that all capital–country associations are injected exclusively via the controlled forward-direction fine-tuning, isolating the reversal curse from pre-existing world knowledge.