# THE ILLUSION OF LATENT GENERALIZATION: BIDIRECTIONALITY AND THE REVERSAL CURSE

**Julian Coda-Forno***
TUM, Helmholtz Munich
julian.coda-forno@helmholtz-munich.de

**Jane X. Wang**
Google DeepMind

**Arslan Chaudhry**
Google DeepMind

## ABSTRACT

The *Reversal Curse* describes a failure of autoregressive language models to retrieve a fact in reverse order (e.g., training on "$A > B$" but failing on "$B < A$"). Recent work shows that objectives with bidirectional supervision (e.g., bidirectional attention or masking-based reconstruction for decoder-only models) can mitigate the reversal curse. We extend this evaluation to include a vanilla masked language modeling (MLM) objective and compare it to decoder-only masking-based training across four reversal benchmarks and then provide a minimal mechanistic study of *how* these objectives succeed. We show that reversal accuracy requires training signal that explicitly makes the source entity a prediction target, and we find little evidence that success corresponds to a single direction-agnostic representation of a fact. Instead, representation distances and linear probes are consistent with storing forward and reverse directions as distinct entries, with different indexing geometry for MLM versus decoder-only masking-based training. Our results caution that objective-level "fixes" can improve reversal behavior without necessarily inducing the kind of latent generalization one might expect from a unified concept.

## 1 INTRODUCTION

Large Language Models (LLMs) can generalize flexibly from information provided *in-context* Lampinen et al. (2025a), yet often fail to use knowledge learned *in the weights* with the same flexibility—a brittleness recently framed as the *latent generalization* problem Lampinen et al. (2025b). A clean symptom is the *Reversal Curse* Berglund et al. (2024): a model trained on a fact in one direction (e.g., "$A > B$") can fail to retrieve the logically equivalent reverse ("$B < A$").

Several mitigation strategies exist. Data-level interventions (e.g., explicit reverse augmentation) can work but may distort language statistics or require manual paraphrase generation Golovneva et al. (2024); Lampinen et al. (2025a). A more fundamental approach modifies the *training objective* to provide bidirectional supervision, either explicitly via bidirectional attention Lv et al. (2024); Nie et al. (2025) or implicitly for decoder-only models via masking-based post-training Pan et al. (2025). These objectives can close the *behavioral* gap, but it remains unclear *how* they do so mechanistically: do they produce a direction-agnostic representation of a fact, or do they instead learn a second, separately-indexed memory entry for the reverse query?

In this paper, we ask what changes *inside* a model when the reversal curse is fixed. We begin by confirming a robust behavioral pattern: objectives that provide bidirectional supervision – explicitly via masked language modeling (**MLM**), or implicitly via masking-based post-training for decoder-only models (**NTP+Masking**) – achieve non-zero reversal accuracy in settings where standard next-token prediction (NTP) collapses (Fig. 1). We then probe what signal is actually doing the work. First, we isolate which prediction targets are necessary for reversal success (Table 1, Fig. 2). Second,

*Work done as a student researcher in GDM.

we analyze the learned representations to test whether success corresponds to a single, direction-agnostic "fact" representation or to storing an additional entry specialized for the reverse query (Figs. 3–4).

Overall, our results point to the latter interpretation: both successful objectives behave as if forward and reverse directions are stored as distinct entries, but the geometry of how these entries are indexed differs between MLM and NTP+Masking.

## 2 SETUP

### 2.1 PROBLEM SETUP & NOTATION

We formalize a fact as a tuple $(s, r, t)$ consisting of a source entity $s$, a relation $r$, and a target entity $t$. In natural language, this is expressed as a sequence $s \xrightarrow{r} t$. Throughout, we use "$A > B$" to denote a forward fact with source $A$ and target $B$, and "$B < A$" to denote its linguistic reverse. Entities $A, B$ may be multi-token spans (e.g., names like *Daphne Barrington*). We always refer to the entity 'A' as source and entity 'B' as target regardless of their position (as subject or target) in the forward or reverse direction.

**Evaluation Protocol (Reversal Accuracy).** For each benchmark, most facts are shown in *both* directions during training to teach that the relation is reversible. For a held-out subset of test facts, the model is trained *only* on the forward direction (e.g., "$A > B$") and evaluated on the reverse prompt (e.g., "$B < \ldots$"). We report the fraction of test prompts (**accuracy**) for which the model outputs the correct source entity (in reverse direction).

### 2.2 DATA

We evaluate on four reversal benchmarks spanning synthetic to semantic structure: **Simple Reversal** Lampinen et al. (2025b), **Nonsense Entities** Lampinen et al. (2025a), **Fictional Celebrity** Berglund et al. (2024), and **Semantic Structure** Lampinen et al. (2025a). Full templates, split ratios, and augmentation details are provided in App. A.

### 2.3 MODELS AND TRAINING

To isolate the effect of finetuning vs pretraining, we use two settings.

**Training from scratch (Simple Reversal).** For *Simple Reversal*, we train models from scratch using the transformer architecture described in Lampinen et al. (2025b). We compare a decoder-only NTP model against an encoder-style MLM variant (no causal mask).

**Pretrained models (language-based datasets).** For *Nonsense Entities*, *Fictional Celebrity*, and *Semantic Structure*, we compare a pretrained BERT-Large (340M) (Devlin et al., 2019) for **MLM** against Gemma-3 4B (Gemma Team et al., 2025) for decoder-only objectives: for decoder-only masking-based training (**NTP+Masking**), we follow Pan et al. (2025) and use the instruct-tuned Gemma-3 4B variant; for standard **NTP**, we use the base Gemma-3 4B.

#### 2.3.1 TRAINING OBJECTIVES

We evaluate three training objectives:

1. **NTP:** standard next-token prediction with a causal mask.
2. **MLM:** masked language modeling with bidirectional attention Devlin et al. (2019).
3. **Masked Fine-Tuning (NTP+Masking).** Following Pan et al. (2025), we train a decoder-only model on examples where a *masked* version of a passage is placed in the context before the *unmasked* passage, and the loss is computed on the unmasked continuation. For a fact "$A > B$", an example is: `[MASK] > B [SEP] A > B`. Intuitively, the masked context encourages the model to infer missing tokens bidirectionally, while still training

with standard next-token prediction on the unmasked segment. At test time, we provide no special tokens and evaluate on the standard reverse prompt "$B <$".

# 3 RESULTS

## 3.1 BIDIRECTIONAL SUPERVISION FIXES THE REVERSAL CURSE

We first establish the baseline performance difference between NTP, MLM, and NTP+Masking across our benchmarks. We evaluate on held-out test sets requiring retrieval in reverse order (Train: "$A > B$", Test: "$B < \ldots$").

Figure 1 replicates the core behavioral claim: providing bidirectional supervision closes the reversal gap. However, this does not yet explain *what* is learned internally—nor whether MLM and NTP+Masking rely on the same mechanism.

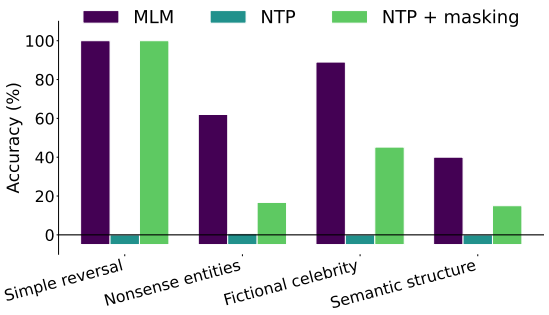


Figure 1: Reversal accuracy across datasets. Standard NTP models fail (near 0%), while MLM and decoder-only masking-based training achieve robust reversal accuracy, indicating that bidirectional supervision (explicit or implicit) is sufficient to mitigate the reversal curse behaviorally.

## 3.2 WHEN DOES MASKING HELP? SOURCE PREDICTION IS NECESSARY (BUT NOT ALWAYS SUFFICIENT)

A reversal query requires predicting the *source* from the *target* plus relation. This motivates a direct test: does reversal require training signal on $p(\text{source} \mid \text{relation}, \text{target})$?

**Ablation: never predict the source.** We modify training so that the **source entity tokens** ($A$) **are never masked**—hence are never predicted. We still allow masking of relation tokens and the target $B$. We then evaluate reversal accuracy as usual. This collapses reversal accuracy to **0%** on all datasets where the ablation is well-defined (Table 1), showing that reversal success depends on ever requiring the model to predict the source given the target and relation.

| Dataset | MLM | NTP+Mask | Abl. I |
|---|---|---|---|
| Simple Reversal | 100% | 99.5% | **0%** |
| Nonsense | 60.2% | 16.7% | **0%** |
| Celebrity | 86.4% | 45.0% | **0%** |

Table 1: Ablation I: removing source prediction (never masking $A$) collapses reversal accuracy to 0% across datasets, despite non-zero reversal accuracy under the standard MLM / NTP+Masking setups (Fig. 1).

**Masking sweep (Simple Reversal).** To isolate what drives success, we focus on *Simple Reversal*—where both MLM and "NTP+Masking" achieve 100% under their standard configurations (Table 1). We then run a *masking sweep*: we vary which components are ever masked (and thus prediction targets)—the *source* $A$, relation token, and/or *target* $B$—and measure reversal accuracy for each variant.

Figure 2 summarizes the sweep. For NTP+Masking (right), making the source $A$ a prediction target is close to sufficient: whenever $A$ is never masked, reversal collapses to 0%, whereas any variant that sometimes masks $A$ yields non-zero performance. For MLM (left), the condition is stricter: masking the source alone does not reliably succeed, and the only consistently successful variant masks **Source & Target**. Thus, even though both objectives require some training signal that targets the source, they differ in how much additional supervision is needed for that signal to generalize to the evaluation query.

## 3.3 REPRESENTATION ANALYSIS: REVERSAL WITHOUT A UNIFIED CONCEPT

Behaviorally, MLM and NTP+Masking both mitigate the reversal curse (Fig. 1). Mechanistically, we ask whether this reflects a direction-agnostic latent concept (forward and reverse co-localized), or two separately-stored entries that merely support the reverse query.

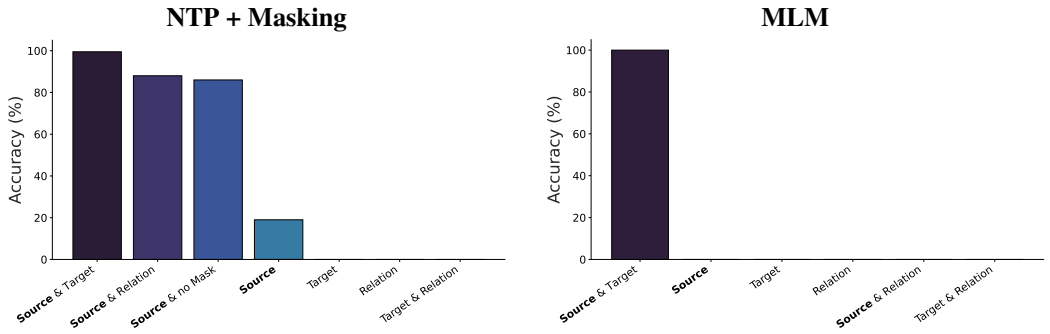


Figure 2: **Ablation: masking sweep on Simple Reversal:** Reversal accuracy as a function of which components are masked (and therefore prediction targets) during training. **Left (NTP + Masking):** masking the *source* is necessary; when the source is never masked, accuracy drops to 0%. **Right (MLM):** masking the source alone is not sufficient; only masking **Source & Target** reliably succeeds.

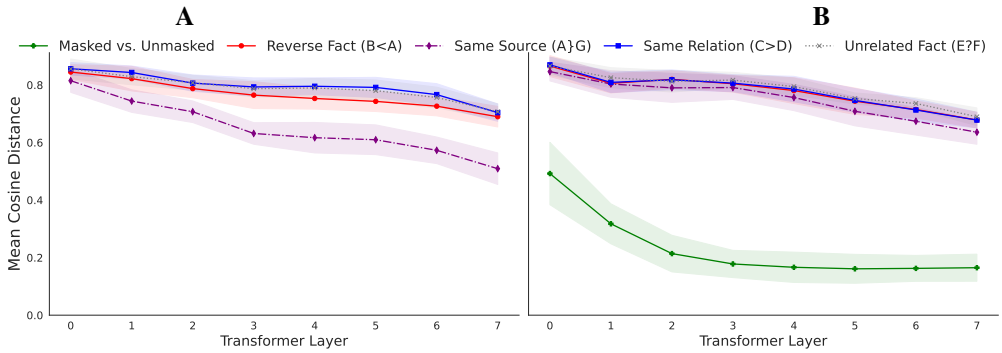


Figure 3: **Representational Distance in MLP Hidden Layers**. Mean Cosine Distance across transformer layers between a fact representation (e.g., $A > B$) and related/unrelated facts for **(A)** NTP + Masking and **(B)** MLM. The reverse fact ($B < A$) remains far from the forward fact, suggesting the two directions are stored as distinct entries rather than a unified direction-agnostic concept.

### 3.3.1 Relative neural distance.

We analyze internal representations within the MLP blocks, which can function as key-value memories Geva et al. (2021). We extract MLP-post-nonlinearity states at the *answer slot*: for NTP/NTP+Masking, the state used to predict the missing entity ("$A > \ldots$"$\rightarrow B$, "$B < \ldots$"$\rightarrow A$); for MLM, the state at `[MASK]` ("$A >$ `[MASK]`" and "$B <$ `[MASK]`" ). MLM provides an internal control by comparing masked vs. unmasked states of the same fact (green curve in Fig. 3B). This "Masked vs. Unmasked" baseline represents the minimum achievable distance for the "same" fact, a reference point not natively available in the decoder-only setup. We measure the *Mean Cosine Distance* between these states and reference points averaged across 20 facts. We use cosine distance specifically because the angular orientation of vectors has been shown to encode stable semantic structures in neural networks Mikolov et al. (2013); Wang et al. (2024).

**Result.** In both objectives, the reverse fact ($B < A$) remains far from the forward representation (Fig. 3), inconsistent with a simple representational "collapse" into a direction-agnostic concept. Instead, both are consistent with "two-slot" storage. Importantly, the detailed structure differs: NTP+Masking shows stronger subject-centric clustering (same-source facts are closer), while MLM shows weaker structure (reverse resembles unrelated facts).

### 3.3.2 LINEAR PROBING FOR INSEPARABILITY.

The significant distances in Fig. 3 suggest that the model's internal geometry does not treat a fact and its reverse as privileged counterparts. However, distance alone does not preclude a more subtle possibility: the two directions might be far apart, yet still be connected by a *consistent mapping*. Concretely, the model might encode a stable transformation that takes the representation of a forward fact to the representation of its reverse—a "semantic bridge" that could implement reversal as a reusable operation rather than as an additional lookup.

This idea is motivated by a classic observation in high-dimensional embedding spaces: relational structure can manifest as consistent offset vectors (e.g., $King - Man \approx Queen - Woman$). If bidirectional objectives truly induce a unified, direction-agnostic understanding of facts, we would expect the forward→reverse relationship to be special in this sense: the vector that connects a fact to its reverse should have a characteristic signature that differs from vectors connecting that fact to arbitrary other memories.

We test this directly using *difference vectors*. For each held-out fact, we compute

$$\Delta_{\text{rev}} \;=\; Fact_1 - ReverseFact_1,$$

and compare it to

$$\Delta_2 \;=\; Fact_1 - Fact_2,$$

where $Fact_2$ is chosen from one of three control sets: *Same Source*, *Same Relation*, or *Unrelated*. We then train a linear probe (logistic regression) to classify whether a given difference vector came from a true reversal pair ($\Delta_{\text{rev}}$) or from a non-reversal comparison ($\Delta_2$).

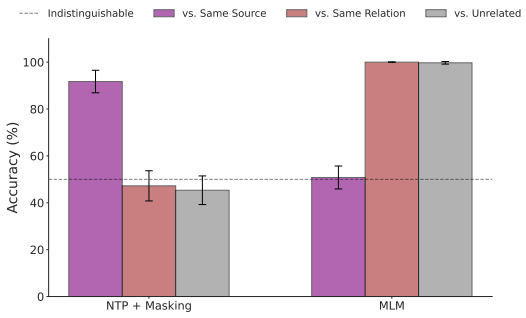


Figure 4: **Linear inseparability of reversals**. Accuracy of a logistic regression probe trained to distinguish $(Fact_1 - ReverseFact_1)$ from $(Fact_1 - Fact_2)$. In NTP+Masking, reversals are indistinguishable from unrelated facts; in MLM, they are primarily indistinguishable from same-source facts.

A key point is how to interpret failure. If the probe achieves only chance accuracy ($\approx 50\%$), then $\Delta_{\text{rev}}$ is *linearly indistinguishable* from $\Delta_2$: the forward→reverse "direction" is no more recognizable to a linear readout than the direction from a fact to some other stored entry. Mechanistically, this means there is no simple, reusable vector-level signature for "reverse this fact" encoded in the representation space. In that case, reversal behavior is more naturally explained by *symmetric encoding*: the reverse direction is encoded as an additional entry, rather than derived via a privileged transformation from the forward one.

Figure 4 shows that this is exactly what happens, but in *different ways* for the two objectives. For NTP+Masking, $\Delta_{\text{rev}}$ is already indistinguishable from *Unrelated* (and *Same Relation*) comparisons: the reverse direction looks, to a linear readout, like a jump to arbitrary knowledge. For MLM, $\Delta_{\text{rev}}$ is primarily indistinguishable from *Same Source* comparisons: reversals are confused with other facts indexed under the same subject. Thus, both objectives solve reversal without an explicit "reverse" transform in representation space, but MLM appears to impose a stronger entity-centric organization than NTP+Masking. This conclusion is specific to *linear* structure and does not rule out a potential systematic *nonlinear* reversal mapping.

## 4 DISCUSSION AND CONCLUSION

Our results separate *behavior* from *mechanism*. Bidirectional supervision (explicit MLM or decoder-only masking-based training) reliably fixes the behavioral reversal curse (Fig. 1), but our analyses suggest this does not correspond to a single direction-agnostic representation of a fact (Figs. 3–4). We do not require exact order-invariant representations: natural-language reversals can differ pragmatically, and inverse relation tokens may be arbitrary. However, if bidirectional supervision induced a unified, direction-agnostic "fact" representation, we would still expect the two directions to become *more* closely coupled in representation space than unrelated facts. Instead of a single direction-agnostic representation, both objectives are consistent with learning two distinct entries

that support forward and reverse queries. Crucially, they organize these entries differently: MLM exhibits stronger subject-centric clustering, while NTP+Masking yields a geometry where reversals can be as indistinguishable as unrelated facts.

**Conclusion.** Mitigating the reversal curse through bidirectional supervision is a meaningful step toward more data-efficient learning, but it may not resolve the underlying brittleness of latent generalization. A direct direction for future work is to design objectives or architectures that *couple* the two directions of a fact—enforcing representational linkage (shared keys, explicit transforms, or consistency constraints) rather than learning a second independently-indexed memory entry.

## REFERENCES

Lukas Berglund, Meg Tong, Max Kaufmann, et al. The reversal curse: Llms trained on "a is b" fail to learn "b is a". *ICLR*, 2024. URL https://arxiv.org/abs/2309.12288.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019. URL https://aclanthology.org/N19-1423/.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories, 2021. URL https://arxiv.org/abs/2012.14913.

Olga Golovneva, Zeyuan Allen-Zhu, Jason Weston, and Sainbayar Sukhbaatar. Reverse training to nurse the reversal curse. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024. URL https://aclanthology.org/2024.emnlp-main.754.pdf.

Andrew K. Lampinen, A. Chaudhry, S. C. Chan, et al. On the generalization of language models from in-context learning and finetuning: a controlled study. *arXiv preprint arXiv:2505.00661*, 2025a. URL https://arxiv.org/abs/2505.00661.

Andrew Kyle Lampinen, Martin Engelcke, Yuxuan Li, Arslan Chaudhry, and James L. McClelland. Latent learning: episodic memory complements parametric learning by enabling flexible reuse of experiences, 2025b. URL https://arxiv.org/abs/2509.16189.

Ang Lv, Kaiyi Zhang, Shufang Xie, Quan Tu, Yuhan Chen, Ji-Rong Wen, and Rui Yan. An analysis and mitigation of the reversal curse. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13603–13615, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.754. URL https://aclanthology.org/2024.emnlp-main.754/.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *ICLR*, 2013.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models, 2025. URL https://arxiv.org/abs/2502.09992.

Xu Pan, Ely Hahami, Jingxuan Fan, Ziqian Xie, and Haim Sompolinsky. Closing the data-efficiency gap between autoregressive and masked diffusion llms, 2025. URL https://arxiv.org/abs/2510.09885.

H. Wang et al. Tracing representation progression: Analyzing and enhancing layer-wise similarity. *arXiv preprint arXiv:2406.14479*, 2024.

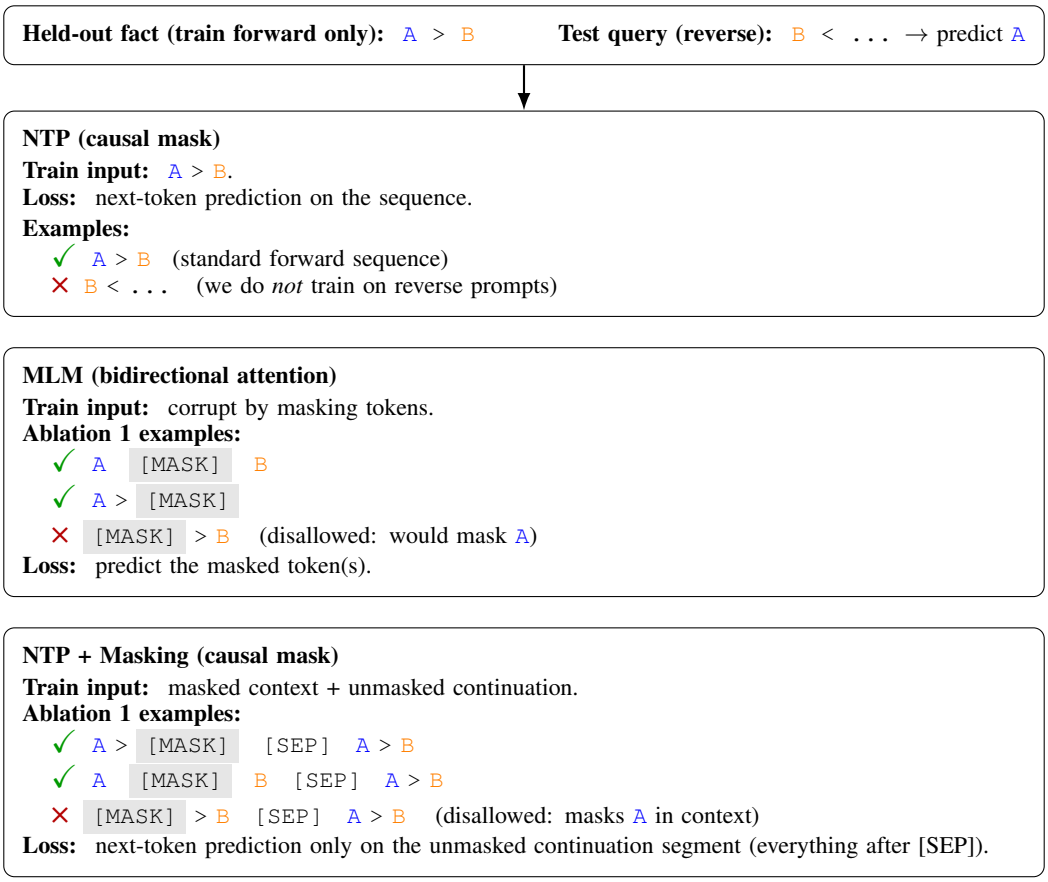


Figure 5: Overview of training setups.

# A Dataset details

## A.1 Datasets

To ensure the robustness of our findings, we utilize four distinct datasets ranging from synthetic to semantic knowledge. Importantly, all datasets are novel, i.e. not contaminated from pre-training.

**Simple Reversal (Lampinen et al., 2025b):** This dataset is explicitly designed to have sufficient diversity for minimal training from scratch. The universe consists of 1,000 entities and 20 relations (plus a reverse relation for each).

- **Data Volume:** The total universe of facts is $1,000$ entities $\times$ 20 relations $\times$ 2 directions $=$ $40,000$ sequences.
- **Split:** We hold out 200 specific facts for testing. For these 200 facts, the model is trained only on the forward sequences. The remaining 39,800 sequences are included in the training set to establish the structural patterns of the relations.
- **Augmentation:** To create diverse training documents, each relation sentence is augmented with up to 5 random prefix and suffix tokens.

**Nonsense Entities (Lampinen et al., 2025a):** This dataset consists of synthetic nonsense words constructed to test structural learning without any potential semantic leakage from pre-training.

- **Example:** "*femp* are more dangerous than *glon*."

- **Structure:** There are 100 such comparisons provided, with comparison words sampled from a set of 28 (e.g., "brighter", "heavier"). Each comparison is repeated across 10 different training examples, paired with randomly sampled preambles.
- **Split:** To facilitate structural learning, 40% of the forward sentences in the training data are explicitly augmented with their reverse counterparts. The test set consists of the remaining facts where the reverse direction was strictly held out.

**Fictional Celebrity (Berglund et al., 2024):** This dataset consists of fictional celebrity names and their descriptions.

- **Example:** "*Daphne Barrington* is the director of *'A Journey through Time'*."
- **Split details:** 50 %.

**Semantic Structure Lampinen et al. (2025a):** This benchmark employs a relational semantic hierarchy with nonsense terms to test deductive inference.

- **Example:** "*gruds*" (dogs) are a type of "*abmes*" (mammals).
- **Structure:** The hierarchy includes 110 categories with naturalistic, asymmetric properties. To mitigate tokenization artifacts, nonsense terms are formed from plausible English phoneme combinations (4-5 letters).

## B Objective details and hyperparameters

**Optimization.** Unless stated otherwise, we fine-tune pretrained models with learning rate $2\times10^{-5}$ for $10^5$ steps. For *Simple Reversal*, we train from scratch using the architecture and training setup in Lampinen et al. (2025b).

**MLM masking rates.** For BERT-style MLM, we use the standard masking procedure: each token is selected for masking with probability $0.15$ Devlin et al. (2019). (For multi-token entities and natural language prompts, this yields diverse corrupted contexts.)

**Decoder-only masking rates (NTP + Masking).** For NTP+Masking on language-based datasets, we follow Pan et al. (2025) and sample a masking ratio uniformly from $(0.05, 0.95)$ per example, then mask tokens accordingly in the context segment placed before the unmasked target segment.

**Special case: Simple Reversal masking.** In *Simple Reversal*, each fact is exactly three tokens ($A$, relation, $B$). Rather than applying a 15% token-level rate (which would rarely mask anything), we construct controlled variants by masking *exactly one* of the three positions (source, relation, or target), or pairs thereof, matching the sweep in Fig. 2.

### B.1 Illustration of training samples on *Simple Reversal*

Table 2 illustrates, for a single fact "$A > B$", what training sequences look like under each objective (for *Simple Reversal*). This is only meant as an intuition pump; for multi-token entities, MLM uses token-level masking with probability 0.15 and NTP+Masking uses a sampled masking ratio.

## C Auxiliary ablations

### C.1 Ablation II: Probing for Representational Collapse

Given that the successful models require predicting $A$ in the context of $B$ to succeed, we investigated the nature of this link. Do these models learn a structured relationship, or does the bidirectional attention (or its approximation) simply cause a "representational collapse" where $A$ and $B$ become strongly associated in a bag-of-words manner, ignoring syntax?

If the representations had collapsed (i.e., the model simply learns "$A$ goes with $B$"), we would expect the model to predict $A$ regardless of the syntactic frame. We tested this by probing the

| **Objective** | **Example training sequences for fact $A > B$** |
|---|---|
| NTP | `A > B` |
| MLM | `[MASK] > B, A [MASK] B, A > [MASK]` |
| NTP + Masking | `A > B, [MASK] > B [SEP] A > B, A [MASK] B [SEP] A > B, A > [MASK] [SEP] A > B` |

Table 2: Illustrative training sequences for *Simple Reversal*. MLM predicts masked tokens directly. NTP+Masking places a masked context before the unmasked continuation and trains the model to reconstruct the unmasked segment.

| **Dataset** | **Training fact example** | **Probe (false frame; relation unchanged)** | **Acc.** |
|---|---|---|---|
| Simple Reversal | $A > B$ | $B >$ [MASK] (would wrongly suggest $A$) | **0%** |
| Nonsense | "femp are **more dangerous than** glon" | "glon are **more dangerous than** [MASK]" (would wrongly suggest femp) | **0%** |

Table 3: Ablation II (syntax probe): if the model had collapsed to a bag-of-entities association, the target ($B$ / "glon") would trigger predicting the training-time source (in red) even in an incorrect syntactic frame. Instead, accuracy remains 0% on these probes.

successful MLM and "NTP + Masking" models with the novel structure: "$B >$ ...?". Note that this is a false statement in our setup (as $B$ is the object), but if the model had merely associated the entities, the presence of $B$ should trigger the retrieval of $A$ even in this incorrect frame. Table 3 lists representative probe instances.

We report this probe for *Simple Reversal* and *Nonsense*. We omit *Celebrity* because that dataset does not train/test on an explicit inverse relational operator: many prompts are effectively definitional ("*X is the director of Y*" $\rightarrow$ "*Who is X?*"), so the "wrong-frame" construction does not cleanly correspond to flipping the syntactic direction while keeping the same relation token. We also omit *Semantic Structure* for the same reason as above: the hierarchical templates make it ambiguous what constitutes a comparable single-slot "false frame" probe.

The models achieved **0% accuracy** on this probe (Table 3). This result indicates that the representations have *not* collapsed. The models distinguish the contexts: they retrieve $A$ when $B$ appears in a reverse frame (or in the "[MASK] $>$" slot), but correctly refuse to associate them in the wrong syntactic direction. Given this, one could argue that MLM is in fact doing latent learning behaviorally. Whether it does so by encoding forward and reverse facts in separate memories is an implementation debate.

### C.2 Ablation III: The Prediction Bias

Finally, we investigated the limits of what information an MLM model actually encodes. In Ablation I, we observed that predicting the source is necessary for reversal. This raises a fundamental question about latent generalization: If a model attends to a token constantly but is never forced to predict it, does it learn that token at all?

To test this, we trained models on the forward sequences "$A > B$". We used a modified MLM objective where we masked $A$ and the relation tokens, but **strictly never masked** $B$. We then tested the model on the *forward* prediction: "$A >$ [MASK]" (expecting $B$). Note that this is not a reversal task; it tests the model's ability to recall the exact sequence seen during training.

Despite $B$ being fully visible in the attention mechanism during every training step (while predicting $A$ or $>$), the model achieved **0% accuracy** when asked to predict $B$ at test time.

This reveals a severe **Prediction Bias**: the model fails to learn representations for tokens that are not explicitly targets of the loss function. Even though $B$ is required to predict $A$, the model treats $B$ merely as a conditioning context and does not encode it as a generatable output. This suggests a bleak outlook for latent generalization in current architectures: information encoded solely in activations (via attention) without a corresponding gradient update on the token itself is effectively invisible to the generation mechanism.

**Connection to post-training.** This offers a lens on a standard post-training choice: many SFT/RLHF pipelines compute loss only on the *answer* tokens, not on the *prompt/question*. This is often desirable (it discourages parroting the prompt), but it may also reinforce the same bias: prompt-side spans can condition behavior without being learned as reliably generatable outputs when those spans are later required.