

Reverse Engineering the Reversal Curse: What Entity Order in Pretraining Does and Does Not Explain

Ivan Samarskyi

Artificial Intelligence Center
Faculty of Electrical Engineering
Czech Technical University
Prague, Czech Republic
samariva@cvut.cz

Ondřej Bojar

Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University
Prague, Czech Republic
bojar@ufal.mff.cuni.cz

Abstract

A causal large language model can retrieve the same fact more confidently in one direction than in the other. This asymmetry is known as the Reversal Curse. The usual explanation links it to next-token prediction: seeing “*A* is *B*” trains the model to predict *B* from *A*, but not *A* from *B*. Prior work supports this with controlled finetuning on synthetic facts. We ask whether the same account still holds in a real pretraining corpus, where facts appear many times, in different contexts, and in both orders. To test this, we work with OLMo 3 7B Base and Dolma, its 5.93-trillion-token stage-1 corpus. For 100 curated relation instances, we locate passages where both entities appear, have two LLM judges check which passages state the expected relation, and probe the model in both directions with contrastive likelihood scores. The raw specific-prompt forward-minus-reverse gap is small and uncertain (0.119, 95% CI $[-0.228, 0.486]$). Once the generic-prompt gap is subtracted, relation naming helps reverse retrieval more than forward ($\Delta^{\text{rel}} = -0.467$, $[-0.673, -0.266]$) on average, though not uniformly across families. That gap is not predicted by local entity order after we control for exposure, proximity, entity frequency, and relation family.

1 Introduction

Ask a model for an author’s debut novel, then ask which author debuted with that novel. Both questions target the same fact, but a causal LLM does not necessarily answer them with equal confidence. This asymmetry—knowing a relation in one direction better than its inverse—is the Reversal Curse (Berglund et al., 2024). Recall depends on whether a fact appears in training, and on how and in which direction it is stored. Figure 1 sketches the idea on a real debut-film pair from our data.

The usual explanation links it to next-token prediction. A sequence that says “*A* [...] *B*” trains the

Pretraining text

A Night of the Living Dead was the debut of **B** George Romero

Ask the model

A → **B** (forward)

Night of the Living Dead

was the debut of _____

George Romero ✓

recall

B → **A** (reverse)

George Romero

debuted with _____

Night of the Living Dead ?

recall

Figure 1: A debut-film pair from our data. The top shows the fact in pretraining text: film (*A*) then director (*B*). The bottom shows the same fact asked in the forward ($A \rightarrow B$) and reverse ($B \rightarrow A$) directions.

model to predict *B* given *A*, not *A* given *B*. Controlled studies support this: researchers write facts in one order, train or fine-tune on them, and find the reverse direction weaker (Kitouni et al., 2024; Golovneva et al., 2024; Lv et al., 2024; Guo et al., 2024), but they do not check the claim against a full pretraining corpus.

A real pretraining corpus is a messier place. A fact is rarely stated just once. It appears in normal text, reviews, metadata, lists, posts, and many other contexts, and while some of these appearances state the expected relation directly, others just put the two entities near each other. The model can also learn a relation from other examples entirely, so wording around a pair in the corpus need not match the wording we use in the probe. Because factual recall tracks pretraining frequency and co-occurrence (Kang and Choi, 2023; He et al., 2025; Liu et al., 2025), we measure how often a pair appears, how close the mentions are, whether the text states the expected relation, and in which order the entities

appear. Without those controls, an apparent order effect could just be frequency or proximity.

We test this with OLMo 3 7B Base (OLMo 3 from now on) and its open Dolma pretraining corpus. We use 100 relation instances in film, games, literature, music, and philosophy. Mappings that are unique in both directions are rare in the world, so we selected pairs in which each side has a single intended partner in our list (§3).

Writing each pair’s entities as X and Y , we scan the stage-1 pretraining corpus, which covers over 97% of all tokens the model saw during training, for documents where both occur, recording which entity appears first and how far apart the two are. Co-occurrence alone does not show that the relation is actually stated, so we sample contexts and ask two LLM judges whether each passage expresses the target relation.

We then probe OLMo 3 in both directions: given a cue X , the model should retrieve Y , and given Y it should retrieve X . We use relation-specific prompts, which name the relation, and generic prompts, which only ask for an association. Generic prompts give a baseline for prompt-direction effects, which we subtract to isolate relation-specific effects.

On every pair, the specific-prompt average is higher for the true cue than for the false-cue baseline, and naming the relation helps further, yet the raw forward–reverse mean sits around zero. Subtracting the generic baseline, we find that naming helps reverse more than forward in the pooled mean. Once exposure, proximity, entity frequency, and family are controlled for, verified corpus order shows little association with that gap.

Our contributions are:

1. a pipeline that connects pretraining co-occurrences, semantically verified relation instances, and retrieval in both directions for the same 100 entity pairs,
2. a relation-specific bidirectional retrieval gap that subtracts the generic bidirectional retrieval gap, and
3. evidence that neither direction is clearly stronger under relation-specific prompts, that the relation-specific gap is negative on average but varies by family, and that after the controls above, measured pretraining order does not reliably predict that gap.

2 Related Work

Reversal and factorization. Berglund et al. (2024) showed that a model trained on synthetic facts in one direction can still fail on the reverse. The Factorization Curse (Kitouni et al., 2024) treats this as one case of the same information under different factorizations. Later work proposes mitigations such as reversed sequences, semantic permutations, and alternative objectives (Golovneva et al., 2024; Guo et al., 2024; Lv et al., 2024). Those studies control the training data and thus isolate order. We instead ask whether, under ordinary pretraining on a large natural corpus, the natural order of entity appearances predicts the direction of later retrieval.

Pretraining exposure and factual recall. Language models encode relational knowledge (Petroni et al., 2019), and retrieval tracks frequency and co-occurrence (Kang and Choi, 2023; Li et al., 2022). The closest work to ours is He et al. (2025), who use earlier OLMo models and Dolma to study how subject and object exposure relate to the recognition of equivalent statements. Unlike that work, we track the order of joint appearances for each pair, separate order from proximity, and keep only contexts that state the expected relation, then test whether that evidence predicts the gap.

Probe measurement. Probes are sensitive to wording (Holtzman et al., 2021; Sclar et al., 2024), and generated answers can diverge from direct probability scores (Hu and Levy, 2023), so we use teacher-forced likelihoods. We average over six templates, compare each true cue to six plausible false cues from the same family, and use generic prompts as a baseline for directional effects unrelated to relation wording.

3 Data and Corpus Evidence

3.1 Relations

We use five relation families with 20 pairs each: film, games, literature, music, and philosophy (Table 1). A bidirectional probe is easier to interpret when each side of a pair has one intended answer, but that property is scarce in the world. A first film can be the debut of more than one director, a studio can ship its first title under a later name, several people can share a first novel or a first album, and a single work can introduce more than one concept. Debut-style relations come closer to uniqueness than most, which is why we use them, but they re-

Code	Relation	n
FIL	first theatrical release → director	20
GAM	debut game → developer studio	20
LIT	debut novel → author	20
MUS	debut album → artist	20
PHI	introducing work → concept	20
Total		100

Table 1: Curated relation families and probe pairs.

main imperfect, and one direction is often cleaner than the other.

For that reason we aim for uniqueness in the pairs we keep, not in the relation types per se. A selected film is treated as that director’s first feature, an album as that artist’s first, and a notion as introduced in that work, and each entity appears in only one pair within its family. False cues come from the other pairs of that family (§4), so among the alternatives the intended answer is unique in both directions.

We generated candidate facts with an LLM, checked them with web search and a second model, and curated the final list by hand. For scanning we fixed a normalized string per entity and used strict alphanumeric, boundary-aware matching. In this way we favour precision at the cost of missing aliases, translations, abbreviations, and pronouns.

3.2 Pretraining scan

We study OLMo 3 and the Dolma 3 mix used for its initial, 5.93-trillion-token pretraining stage (Ettinger et al., 2025). The later 100B-token mid-training stage and 50B-token long-context stage are much smaller, so the stage-1 corpus that we scan accounts for over 97% of the full training data, and all corpus claims in this paper refer to stage-1 only.

We scan the corpus with an Aho–Corasick automaton (Aho and Corasick, 1975) over normalized entity forms. For each pair–document training instance, we keep the nearest non-overlapping mention pair, record which entity comes first, and compute the character distance in normalized text. Dolma can repeat source documents through mixture weighting, so these counts reflect training instances rather than unique documents.

The scan finds 2,370,008 pair–document training instances, of which 1,804,582 have nearest mentions within 150 characters. We set the window at 150 characters so that the two mentions fall within roughly one sentence of each other, which is where

a passage that states the relation would normally place them. Verification then expands the surrounding context to sentence boundaries before judging, so the judge sees more than the window itself. Proximity alone does not mean the relation is stated, so we send these co-occurrences to semantic verification.

From the raw co-occurrences, we compute a proxy order score:

$$r_i^{\text{proxy}} = \log_2 \frac{C_i^{X<Y} + 1}{C_i^{Y<X} + 1}. \quad (1)$$

Here $C_i^{X<Y}$ is the number of pair–document training instances where X appears before Y , and $C_i^{Y<X}$ is the number where Y appears before X . We keep this score separate from total co-occurrence and distance so that order can be tested apart from exposure and proximity.

3.3 Semantic verification

Proximity is not evidence that the expected relation is present, because a film and a director can sit near each other in a cast list or a paragraph about something else. Such a passage may never say that the director made the film, and it rarely says that the film was the director’s debut. That is why we check whether each sampled context actually states the intended relation.

Identical occurrences of the same pair in the same source document are judged only once, but we keep their training-mixture multiplicity. We sample unique contexts within strata defined by mention order and distance band, capped at 100 per pair, keeping each stratum’s inclusion probability so that every judgment can be weighted by both inverse inclusion probability and mixture multiplicity. Two LLM judges, GPT-OSS-120B and Qwen3.5-122B, label each sampled passage as stating the relation, not stating it, or uncertain. The prompt (Appendix A) does not include the computed mention-order label, although order remains visible in the passage. For every positive label, we require a grounding phrase from the original text, and disagreements go to a separate LLM adjudicator, DeepSeek-V4-Pro. Invalid or ungrounded positive outputs are not counted as relation evidence.

After reweighting, we obtain evidence totals $D_i^{X<Y}$ and $D_i^{Y<X}$ for each pair. These totals define the verified order score:

$$r_i^{\text{ver}} = \log_2 \frac{D_i^{X<Y} + 1}{D_i^{Y<X} + 1}. \quad (2)$$

We use this as the main order predictor. For eight pairs, both directions have zero accepted evidence, so the smoothed ratio is zero but uninformative. We drop those pairs from regressions on r_i^{ver} while keeping them in probe summaries and the raw-order robustness check.

In total, we judged 9,646 unique contexts, of which 1,414 received a final entailed label. The mean pair-level judge agreement rate was 89%, the adjudicator was used for 11% of contexts, and 0.2% of judge responses were invalid. After sampling and multiplicity weighting, the mean pair-level acceptance rate was 11.4%, and pooling over all 1,804,582 within-window training instances gives about 9.7%. Roughly one nearby co-occurrence in ten states the expected relation, so we treat raw co-occurrence as a noisy proxy and use verified relations for the main test.

4 Contrastive Retrieval Measurement

For each pair (X_i, Y_i) , the forward probe uses X_i as the cue and Y_i as the target, while the reverse probe swaps them. In the film family, for example, forward means work-to-director and reverse means director-to-work, and we test each direction with two prompt styles. *Specific* prompts name the relation, such as “__ was the first feature film directed by”. *Generic* prompts ask for a connection without naming the relation, for example “__ is regularly linked with”. Each condition uses six templates (Appendix A).

For each template, we compare the log-probability of the true target under the true cue with the log-probability of the same target under six false cues from the same relation family and the same cue role: other film titles when the cue is a film, other creators when the cue is a creator. The target string and template wording are fixed, and only the cue entity changes:

$$\ell_{ik} = m(T_i | P_{ik}^{\text{true}}) - \frac{1}{6} \sum_{j=1}^6 m(T_i | P_{ijk}^{\text{false}}). \quad (3)$$

Here m is the teacher-forced mean log-probability per target token, and total target log-probability is kept as a robustness specification. The main score measures how much the cue helps retrieve the target, rather than how common the target is alone, and the false-cue baseline is a within-family contrast for the same target rather than a corpus-frequency baseline.

Averaging over templates gives four lift values per pair, $\ell_i^{s,f}, \ell_i^{s,r}, \ell_i^{g,f}, \ell_i^{g,r}$, where s/g denote specific or generic prompts and f/r denote forward or reverse direction. This gives $2 \times 2 \times 6 \times 7 = 168$ scored sequences per pair, or 16,800 sequences in total.

We define the mean specific and generic lift as

$$L_i^s = \frac{1}{2}(\ell_i^{s,f} + \ell_i^{s,r}), \quad L_i^g = \frac{1}{2}(\ell_i^{g,f} + \ell_i^{g,r}). \quad (4)$$

Their difference, $S_i = L_i^s - L_i^g$, measures how much naming the relation adds beyond a generic connection.

For bidirectional retrieval, we first compute the signed specific and generic bidirectional retrieval gaps, $\Delta_i^s = \ell_i^{s,f} - \ell_i^{s,r}$ and $\Delta_i^g = \ell_i^{g,f} - \ell_i^{g,r}$. We use the generic gap as a baseline for the directional differences that remain when the relation is not named, including differences in wording, entity role, and target length. Our main outcome, the *relation-specific bidirectional retrieval gap*, subtracts this baseline:

$$\Delta_i^{\text{rel}} = \Delta_i^s - \Delta_i^g. \quad (5)$$

A positive Δ_i^{rel} means naming the relation helps forward more than reverse, relative to the generic baseline, and a negative value means the opposite. This contrast alone does not say which direction is stronger in raw terms, which depends on Δ_i^s .

5 Experimental Setup

We score OLMo 3 7B with batched teacher forcing. The main regression predicts Δ^{rel} from verified order while controlling for other factors:

$$\Delta_i^{\text{rel}} \sim r_i^{\text{ver}} + \log(1 + C_i) + q_i + \log_2 \frac{f_{X_i} + 1}{f_{Y_i} + 1} + \text{family}_i, \quad (6)$$

where C_i is total pair co-occurrence, q_i is the share of co-occurrences within 100 characters, and f_X, f_Y are marginal document frequencies for the two entities. All continuous predictors are standardized, so the coefficient on r_i^{ver} is the expected change in Δ^{rel} for a one-standard-deviation increase in verified order. Family is controlled with Philosophy (PHI) as the reference level, and using a different reference family would not change this coefficient.

Regression coefficients are estimated by OLS, and descriptive summaries are pair means. Confidence intervals for the main regression and for

Analysis	Quantity	Est.	95% CI	<i>n</i>
Probe	Specific mean lift	2.499	[2.331, 2.655]	100
Baseline	Generic mean lift	1.628	[1.468, 1.786]	100
Probe	Specificity gap	0.870	[0.767, 0.978]	100
Direction	Specific bidirectional gap Δ^s	0.119	[-0.228, 0.486]	100
Direction	Relation-specific bidirectional gap Δ^{rel}	-0.467	[-0.673, -0.266]	100
Main	Verified order $\rightarrow \Delta^{\text{rel}}$	-0.043	[-0.235, 0.149]	92
Raw outcome	Verified order $\rightarrow \Delta^s$	-0.163	[-0.525, 0.199]	92
Total log-prob.	Verified order $\rightarrow \Delta^{\text{rel}}$	0.123	[-0.439, 0.665]	92
Proxy order	Proxy order $\rightarrow \Delta^{\text{rel}}$	0.010	[-0.355, 0.376]	100
Quality	Verified order, effective $n \geq 20$	-0.049	[-0.229, 0.130]	89
Exploratory	Proximity $\rightarrow$ generic lift	0.372	[0.099, 0.645]	100

Table 2: Main estimates. Mean, main-regression, and total-log-probability intervals use family-stratified bootstrap. The other regression intervals use HC3 standard errors. Regression slopes are per one standard deviation of the named predictor. Verified-order analyses exclude pairs with no accepted relation evidence.

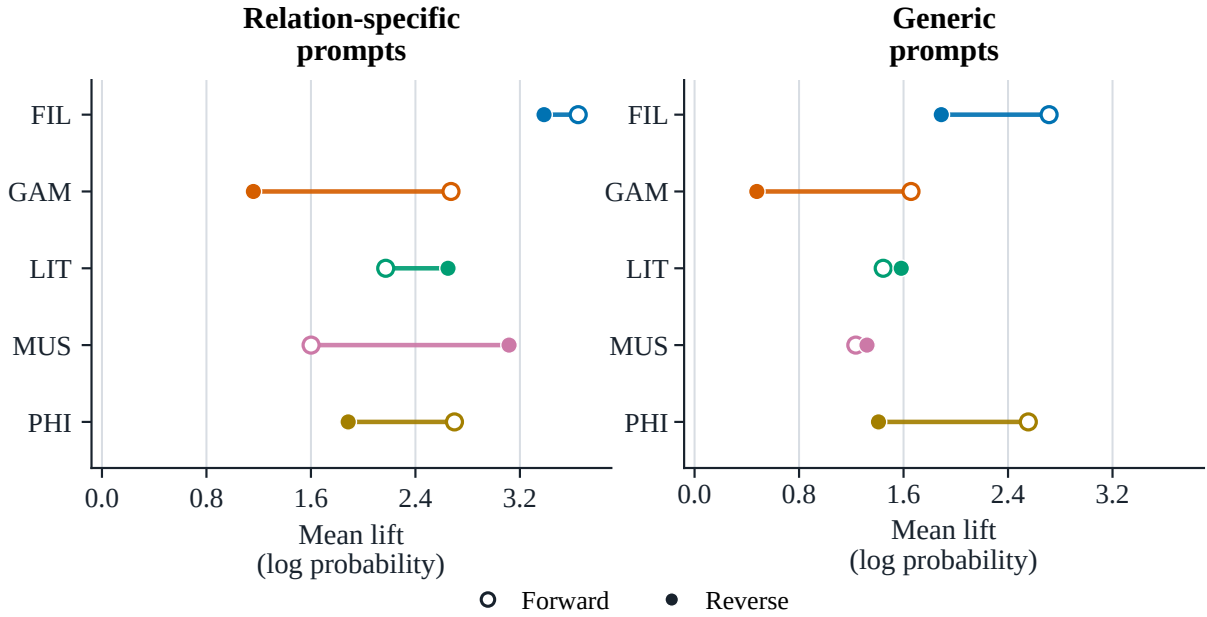


Figure 2: Forward ($X \rightarrow Y$) and reverse ($Y \rightarrow X$) mean lift under relation-specific and generic templates, broken down by relation family. Open markers indicate forward retrieval, filled markers indicate reverse retrieval, and connecting segments show the within-family directional difference.

the mean estimates both use 5,000 family-stratified bootstrap resamples of pairs.

We also rerun the analysis with raw proxy order, total target log-probability, the unsubtracted specific-prompt gap, other proximity windows, and leave-one-family-out models.

6 Results

6.1 True cues identify the intended pairs

With specific prompts, the true cue gives a higher pair-level mean than the false-cue baseline for all 100 pairs, although four pairs have one individual direction below the baseline (Table 2, Figure 3). The true cue raises the target mean per-

token log-probability by 2.499 relative to the six false cues (95% CI [2.331, 2.655]). These scores are natural log-probabilities, so the average corresponds to $\exp(2.499) \approx 12.2$ times the geometric mean per-token likelihood of the target compared with the same-family false-cue average. Generic prompts also show a strong association (1.628, [1.468, 1.786]), or $\exp(1.628) \approx 5.1$ on the same scale, and naming the relation adds another 0.870 ([0.767, 0.978]), about a further $\exp(0.870) \approx 2.39$ -fold gain, so the intended pair is well separated from same-family alternatives.

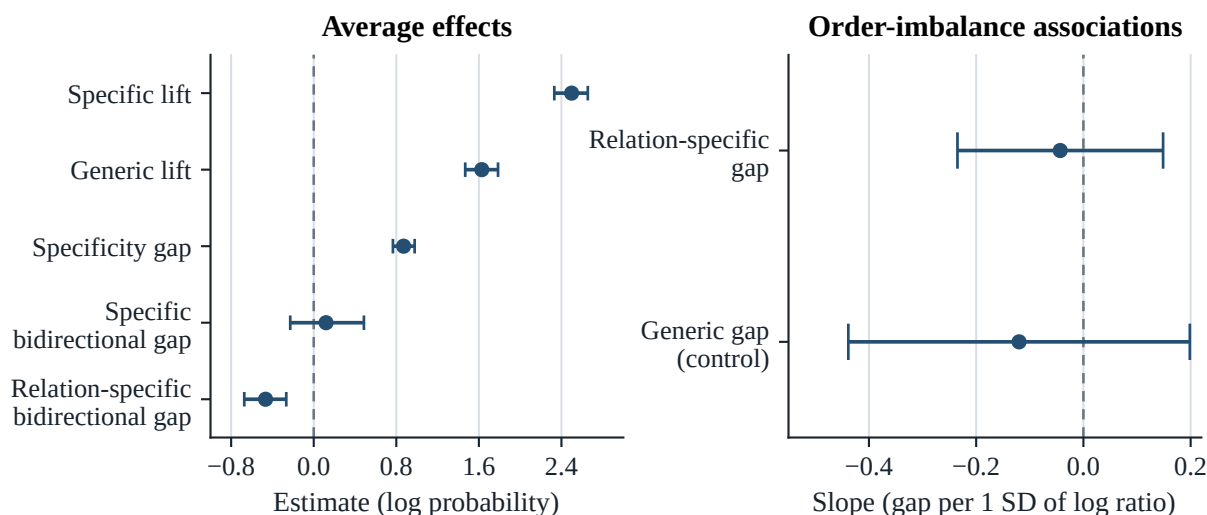


Figure 3: Summary estimates for the main analyses. Left: average contrastive retrieval effects, including the specific and relation-specific bidirectional retrieval gaps. Right: order-imbalance associations with the relation-specific and generic bidirectional retrieval gaps. Error bars are 95% confidence intervals.

6.2 Relation wording changes the bidirectional retrieval gap

Across all 100 pairs, the specific-prompt bidirectional retrieval gap, forward minus reverse, is 0.119 (95% CI $[-0.228, 0.486]$), with mean forward and reverse lifts 2.558 and 2.439, so these prompts give no evidence of an overall reverse retrieval advantage.

Generic prompts favour forward retrieval by 0.586 on average. After subtracting that baseline, the mean relation-specific gap is $\Delta^{\text{rel}} = -0.467$ ($[-0.673, -0.266]$): naming the relation adds 0.637 over the generic baseline in the forward direction and 1.104 in the reverse direction. Word-ing helps reverse more in the pooled mean, without implying that reverse retrieval is stronger in raw terms. Figure 2 shows the underlying forward and reverse lifts under both prompt types.

The gap varies by family (Figure 4): the means are -0.564 for Film ($[-0.918, -0.227]$), 0.333 for Games ($[-0.258, 0.893]$), -0.339 for Literature ($[-0.740, 0.079]$), -1.430 for Music ($[-2.009, -0.881]$), and -0.334 for Philosophy ($[-0.725, 0.067]$). Only Film and Music have family intervals wholly below zero, so we avoid stronger claims about the remaining families.

6.3 Appearance order is not a reliable predictor of the gap

If local order drove relation-specific directional behavior, pairs seen more often as X before Y should show a more positive Δ^{rel} . Among the 92

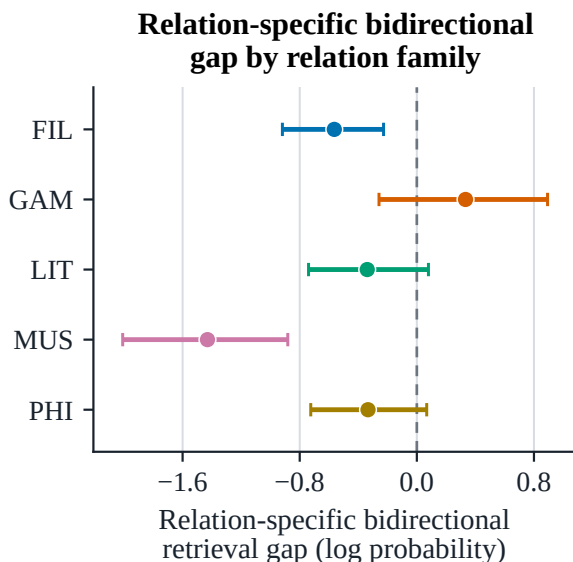


Figure 4: Relation-specific bidirectional retrieval gap by relation family, with 95% bootstrap confidence intervals. Negative values mean that relation naming helps reverse retrieval more than forward retrieval, relative to the generic baseline.

pairs with accepted relation evidence, the standardized order coefficient is instead -0.043 (95% CI $[-0.235, 0.149]$), a small estimate whose interval crosses zero (Figure 5, see also Figure 3). The interval still allows a small effect either way: this is not evidence that the true effect is exactly zero, only that measured local order is not a reliable predictor of the gap in this setting.

The same pattern holds across outcomes and evidence choices. Total target log-probability gives

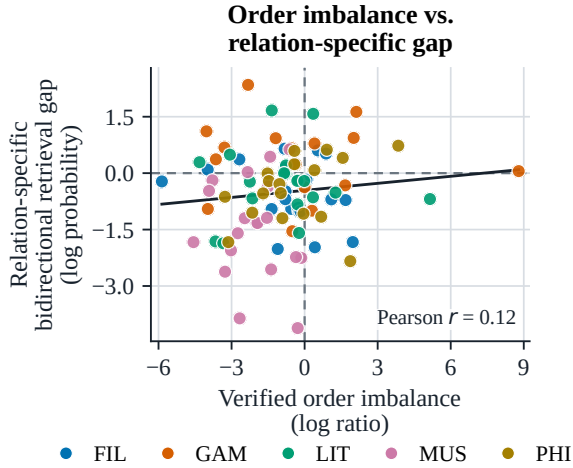


Figure 5: Verified order score versus the relation-specific bidirectional retrieval gap for the 92 pairs with accepted evidence. Pearson r is descriptive. The main claim uses the controlled regression in Equation 6.

0.123 $([-0.439, 0.665])$, and verified order also fails to predict the unsubtracted specific-prompt gap $(-0.163, [-0.525, 0.199])$. Raw proxy order on all 100 pairs gives 0.010 $([-0.355, 0.376])$, and requiring an effective sample size of at least 20 gives -0.049 $([-0.229, 0.130], n = 89)$. Replacing the proximity control with the share of pairs within 50 or 150 characters leaves the slope unchanged $(-0.042$ and $-0.043)$. Leave-one-family-out coefficients range from -0.074 to 0.003 , all with intervals crossing zero, and verified order does not predict the generic-prompt gap either $(-0.120, [-0.438, 0.198])$. Nine of the 92 pairs have accepted evidence in only one direction, and excluding them gives -0.045 $([-0.342, 0.253], n = 83)$.

6.4 Proximity is associated with generic retrieval

In an exploratory analysis, the share of co-occurrences within 100 characters predicts generic association lift after controlling for marginal frequency and family ($\hat{\beta} = 0.372$, 95% CI $[0.099, 0.645]$). We treat this as descriptive: closeness in the corpus tracks how strongly the model links the entities, while measured local order shows no reliable link to either bidirectional gap.

7 Discussion

True cues identify the intended targets and naming the relation improves retrieval, so the facts are in the model. What remains is directional: the raw specific-prompt forward–reverse mean is compati-

ble with zero, and relative to the generic baseline relation wording still helps reverse more than forward. Neither raw nor verified order predicts that gap once exposure, proximity, entity frequency, and family are controlled for, and verified order also fails to predict the raw specific-prompt gap.

A single next-token example is still directional: “A ... B” supervises later tokens from earlier ones, but pretraining is not one controlled sequence. Familiar pairs recur across sources and phrasings, and our order score only records which entity’s nearest mention comes first inside a local window.

Measurement error in the order score can shrink the estimated association toward zero. Verified totals come from stratified samples, yet the regression treats the resulting ratios as fixed covariates, so intervals may understate uncertainty, and missed true mentions add further noise. Normalized scan forms and the verified subset cut down on false positives, but we still miss aliases and any relation evidence outside the 150-character window. The same weak association shows up with proxy order, a stricter evidence filter, and leave-one-family-out checks, so it is unlikely to be an artifact of one family or one evidence threshold alone.

The gap is clearly negative for Film and Music, while the intervals for the other three families cross zero. Family variation may reflect entity type, cue specificity, or syntax, but our experiments do not distinguish among these. That pattern is hard to reconcile with local appearance order as the sole explanation.

8 Conclusion

For these OLMo 3 facts, local entity order in pretraining does not predict directional retrieval. Relation-specific prompts improve over generic ones, and the subtracted bidirectional gap varies by family, but mention order does not track that gap even after semantic verification and the checks above. Next-token prediction makes individual training examples directional, yet the local mention order we measure is not a reliable explanation for the gap in these facts. Closing the remaining gap would need other measurements (entity type, cue specificity, richer context, or training dynamics), or a direct intervention on the training data.

Limitations

One model and incomplete training history. We study a single 7B base model, and results may

differ for other model sizes, architectures, or a different tokenizer. The stage-1 corpus we scan accounts for over 97% of all tokens the model saw during training, but we do not scan the later stages, which may contain relation evidence that is not measured here.

Scan forms and local contexts. The scan uses exact normalized forms, which keeps matches precise, but it misses aliases, translations, abbreviations, pronouns, and other non-canonical references. Semantic verification is also limited to contexts where the two appearances are within 150 characters, so the order score we measure is not exhaustive.

LLM-based verification. The verification step uses LLM judges and an adjudicator on a subset of sampled contexts, at most 100 per pair. We did not compare the judges with human annotations, so systematic judging errors remain possible.

Observational approach. We only work with the order already present in the corpus and we do not retrain the model with that order flipped. So the fact that measured order does not reliably predict the measured gaps does not mean that changing order during training would do nothing. Our order score may also miss the order the model actually learns from: it records which name comes first in the nearest mention pair of a document, while the model sees every mention, including aliases and pronouns, and also learns from the surrounding words. Tracking that at corpus scale is beyond this study.

Scope of probing. We use six templates per condition, but prompt phrasing can still matter. The forward and reverse templates are semantically paired but cannot be made identical. Relations that are unique in both directions in the world are scarce, for example, debut credits can be shared, studio names change, and a work can introduce more than one idea. We selected pairs so that each cue has one intended answer in our list. Family-level gaps should not be perceived as if each relation type were a perfect bijective mapping. The results may also not generalize to other relation types.

Acknowledgments

This work was supported and computational resources were provided by OP VVV project CZ.02.1.01/0.0/0.0/16 019/0000765 RCI. Access to open-source LLMs and computational resources

were also provided by the e-INFRA CZ project, supported by the Ministry of Education, Youth and Sports of the Czech Republic.

Ondřej Bojar would like to acknowledge the support from the grant CZ.02.01.01/00/23_020/0008518 (“Jazyková, umělá inteligence a jazykové a řečové technologie: od výzkumu k aplikacím”).

References

- Alfred V. Aho and Margaret J. Corasick. 1975. [Efficient string matching: An aid to bibliographic search](#). *Communications of the ACM*, 18(6):333–340.
- Lukas Berglund, Meg Tong, Maximilian Kaufmann, Mikita Balesni, Asa Stickland, Tomek Korbak, and Owain Evans. 2024. [The reversal curse: LLMs trained on “A is B” fail to learn “B is A”](#). In *International Conference on Learning Representations*, volume 2024, pages 18623–18642.
- Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, and 48 others. 2025. [Olmo 3](#). *Preprint*, arXiv:2512.13961.
- Olga Golovneva, Zeyuan Allen-Zhu, Jason Weston, and Sainbayar Sukhbaatar. 2024. [Reverse training to nurse the reversal curse](#). *arXiv preprint arXiv:2403.13799*.
- Qingyan Guo, Rui Wang, Junliang Guo, Xu Tan, Jiang Bian, and Yujiu Yang. 2024. [Mitigating reversal curse in large language models via semantic-aware permutation training](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11453–11464.
- Yuan He, Bailan He, Zifeng Ding, Alisia Lupidi, Yuqicheng Zhu, Shuo Chen, Caiqi Zhang, Jiaoyan Chen, Yunpu Ma, Volker Tresp, and Ian Horrocks. 2025. [Supposedly equivalent facts that aren’t? entity frequency in pre-training induces asymmetry in LLMs](#). *arXiv preprint arXiv:2503.22362*. Version 2.
- Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. [Surface form competition: Why the highest probability answer isn’t always right](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7038–7051.
- Jennifer Hu and Roger Levy. 2023. [Prompting is not a substitute for probability measurements in large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, Singapore. Association for Computational Linguistics.

Cheongwoong Kang and Jaesik Choi. 2023. [Impact of co-occurrence on factual knowledge of large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7721–7735.

Ouail Kitouni, Niklas Nolte, Diane Bouchacourt, Adina Williams, Mike Rabbat, and Mark Ibrahim. 2024. [The factorization curse: Which tokens you predict underlie the reversal curse and more](#). *arXiv preprint arXiv:2406.05183*.

Shaobo Li, Xiaoguang Li, Lifeng Shang, Zhenhua Dong, Chengjie Sun, Bingquan Liu, Zhenzhou Ji, Xin Jiang, and Qun Liu. 2022. [How pre-trained language models capture factual knowledge? a causal-inspired analysis](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1720–1732.

Yihong Liu, Mingyang Wang, Amir Hossein Kargaran, Felicia Körner, Ercong Nie, Barbara Plank, François Yvon, and Hinrich Schuetze. 2025. [Tracing multi-lingual factual knowledge acquisition in pretraining](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 2121–2146.

Ang Lv, Kaiyi Zhang, Shufang Xie, Quan Tu, Yuhao Chen, Ji-Rong Wen, and Rui Yan. 2024. [An analysis and mitigation of the reversal curse](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13603–13615.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. [Language models as knowledge bases?](#) In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. [Quantifying language models’ sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting](#). In *The Twelfth International Conference on Learning Representations*.

A Verification Prompts and Cue Templates

Table 3 lists the cue templates (§4), with [CUE] replaced by the cue entity. Below are the exact prompts used for semantic verification (§3). GPT-OSS-120B and Qwen3.5-122B receive the judge prompts, DeepSeek-V4-Pro receives the adjudicator prompts together with both judgments, and is called only when they disagree on the label or the relation type. We keep an entailed label only if the returned grounding phrase appears verbatim in the passage (at most 20 words).

Judge: system prompt

You are a factual-relation classifier.

Determine whether the passage itself states or grammatically entails the canonical fact $X \{relationship\} Y$. The passage must support the complete relation, not merely a broad association; include role-specific wording such as “debut”, “first”, or “introduced” when those words are part of the fact. If the fact is entailed, `grounding_phrase` must be a short verbatim contiguous quote from the passage, not a paraphrase. Prefer the smallest span that still proves the relation.

Use only the passage. Do not use outside knowledge. Mention order does not determine the answer. Active voice, passive voice, possessives, and appositives may express the same relation when the semantic roles remain X and Y .

Return valid JSON only with this exact shape: `{"relation_label": "entailed" | "not_entailed" | "uncertain", "relation_type": "explicit" | "implicit" | "none", "grounding_phrase": "..."}"`

Judge: user prompt

Entities: $X = \{x\}$, $Y = \{y\}$

Canonical fact: $X \{relationship\} Y$

Passage: `{context}`

Decide whether the passage states or grammatically entails the canonical fact. The passage must support the complete relationship, not just loose association; it must explicitly or grammatically entail the full claim, including role words such as “debut”, “first”, or “introduced” when they matter to the relation.

If the fact is entailed, the `grounding_phrase` must be a short contiguous quote from the passage, copied verbatim, kept as short as possible while still proving the claim, and present in the passage after normalization. Prefer the smallest span that makes the relation obvious. If no short contiguous span supports the claim, return “uncertain” or “not_entailed” rather than inventing a paraphrase. Return JSON only.

Adjudicator: system prompt

You resolve disagreements between factual-relation judges. Use only the passage and the supplied judgments. If you return “entailed”, `grounding_phrase` must be a short verbatim contiguous quote from the passage. Keep it minimal and do not paraphrase. Return valid JSON only with keys: `{"relation_label": "entailed" | "not_entailed" | "uncertain", "relation_type": "explicit" | "implicit" | "none", "grounding_phrase": "..."}"`

Adjudicator: user prompt

Context: `{context}`

Entity X : `{x}`

Entity Y : `{y}`

Canonical fact: $X \{relationship\} Y$

`{judgments}`

Resolve the disagreement and return one JSON object only. If the final answer is entailed, `grounding_phrase` must be a short contiguous quote from the context, copied verbatim and kept as short as possible.

Forward cue (work → creator)	Reverse cue (creator → work)
<i>Film</i>	
[CUE] was the first feature film directed by	[CUE] made their feature-directing debut with
The director whose first feature was [CUE] is	The first feature directed by [CUE] was
[CUE] marked the feature-film directing debut of	The feature-film directing debut of [CUE] was
[CUE] was the debut feature of director	The debut feature of [CUE] was
The filmmaker who debuted as a feature director with [CUE] was	[CUE] debuted as a feature director with
[CUE] launched the feature-directing career of	The feature-directing career of [CUE] began with
<i>Games</i>	
[CUE] was the debut video game developed by	[CUE] made its game-development debut with
The developer that released [CUE] as its first game was	The first game released by [CUE] was
[CUE] marked the game-development debut of	The game-development debut of [CUE] was
[CUE] was the inaugural game project of	The inaugural game project of [CUE] was
The studio whose debut game was [CUE] is	The debut game from [CUE] was
[CUE] launched the game-development catalog of	[CUE] launched its game-development catalog with
<i>Literature</i>	
[CUE] was the debut novel of	[CUE] made their literary debut with
The author whose first novel was [CUE] is	The first novel by [CUE] was
[CUE] marked the novel-writing debut of	The novel-writing debut of [CUE] was
[CUE] was the first published novel by	The first published novel by [CUE] was
The writing career launched by [CUE] belonged to	[CUE] launched their writing career with
[CUE] was the maiden novel of	The maiden novel by [CUE] was
<i>Music</i>	
[CUE] was the debut album of	[CUE] made their recording debut with
The artist whose first album was [CUE] is	The first album by [CUE] was
[CUE] marked the recording debut of	The recording debut of [CUE] was
[CUE] opened the discography of	[CUE] opened their discography with
The musical career launched by [CUE] belonged to	[CUE] launched their musical career with
[CUE] was the inaugural album released by	The inaugural album released by [CUE] was
<i>Philosophy (work → concept)</i>	
[CUE] was the work that first introduced the notion of	[CUE] was first introduced in
The concept originally introduced in [CUE] was	The work that originally introduced [CUE] was
[CUE] contained the first formulation of	The first formulation of [CUE] appeared in
[CUE] was the introducing work for	The work that introduced [CUE] was
Scholars credit [CUE] with first defining	Scholars trace the first definition of [CUE] to
[CUE] presented the earliest account of	The earliest account of [CUE] appeared in
<i>Generic association (all families, both directions)</i>	
[CUE] is often mentioned alongside	
[CUE] is frequently discussed together with	
[CUE] is commonly associated with	
[CUE] is regularly linked with	
Discussions of [CUE] often also mention	
A name frequently appearing with [CUE] is	

Table 3: Cue templates for the contrastive probe. Each row is a forward/reverse pair (six per family). Generic templates are shared across families and directions.