Life of Research · Published 19 November 2026 · DOI 10.5281/zenodo.24372087

A Learned Inverse Is Not a Lifted Curse: What Training Transformers on a Family Knowledge Graph Shows About the Reversal Curse, Why Its Bilinear Geometry Answers a Relation-Level Question the Original Curse Did Not Ask, and the Controls That Would Carry It to Pretrained Models

Pranay Mahendrakar
SONYTECH · ORCID 0009-0003-7224-029X
pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. A 2025 study trained decoder-only transformers from scratch on a synthetic family knowledge graph and reported that, under strong weight decay, some models infer withheld parent relations from the child relations they did see, that a bilinear probe in the style of the RESCAL knowledge-graph model reads their middle layers at above 95 percent, and that the same models propagate a single fine-tuned edit to its logical consequences while models without the structure do not. Its authors conclude that the reversal curse is an artifact of how knowledge is encoded rather than an inherent limitation. In this analysis the author reviews and synthesises published results to show that the experiment is real and that it answers a different question from the one the reversal curse posed. The curse, as first reported, concerns a single fact whose reverse appears nowhere and one of whose terms is not an entity the model represents in any other role; the anchor's test asks for an inverse relation between two entities that each appear in four to eight training facts, each repeated about 20,000 times, after the same inverse pair has been demonstrated on 5,000 other entities. Knowledge-graph completion learned years ago that this is the easiest pattern in relational data: a rule that only reads inverse pairs reached 96.4 percent filtered Hits@1 on WN18. What remains genuinely contested is why next-token transformers learn that relation-level rule in some from-scratch studies and fail almost completely in others with the same design logic, and the published differences include the regularisation strength the anchor varies. The paper derives that a bilinear score yields the inverse by transposition only when subject and object are read through the same encoding, a property the anchor's probe assumes rather than tests. It argues that the probe-to-editing correlation (R squared 0.939) does not separate geometry from the generalising solution that produces both. It consolidates 25 published results in a table and a figure, states a procedure for reading reversal and propagation claims, and names six measurements that would settle the open part.

Keywords: reversal curse; knowledge editing; relational representation; bilinear models; knowledge graph completion; grokking; probing; language models

1. Introduction

A language model fine-tuned on "Valentina Tereshkova was the first woman to travel to space" will answer who Tereshkova was and will not, without further help, answer who the first woman in space was. The original demonstration found accuracy in the reverse direction close to zero for fictitious facts and a likelihood for the correct name no higher than for a random one (Berglund et al., 2023). Influence functions computed on models of up to 52 billion parameters show the same asymmetry from the training side: a training sequence loses almost all its influence on a query once its key phrases appear in the opposite order (Grosse et al., 2023). Pretraining from scratch on synthetic biographies finds that searching for a person from an attribute stays at virtually zero unless the knowledge appears in reversed form in the data (Allen-Zhu & Li, 2023b). This failure acquired a name, the reversal curse, and a reputation as a structural property of autoregressive training. Theoretical and empirical work implicates the next-token objective and its training dynamics (Zhu et al., 2024; Lv et al., 2023), and a reframing as a factorization curse, the failure to learn one joint distribution under different orderings, reports that it survives scaling, token reversal and simple bidirectional attention (Kitouni et al., 2024).

Against that reputation, Kim et al. (2025) report that the curse can be lifted without changing the objective. They train 206-million-parameter GPT-NeoX models from scratch on a synthetic knowledge graph of 1,000 families, withhold the father and mother relations for half of the families, and test whether the models infer them. With weak regularisation every model fails. With strong weight decay some models succeed almost perfectly, and in those models a bilinear probe, which scores a pair of entity vectors through a relation-specific matrix as the RESCAL model of knowledge-graph embedding does (Nickel et al., 2015), reads relations from the middle layers with high accuracy, while the transpose of a relation matrix behaves like its inverse and the product of two matrices like their composition. When one fact in such a model is edited by fine-tuning, the change propagates to the reverse fact and to the parent relations it entails; in models without the structure it does not. The authors draw two conclusions:

© 2026 Pranay Mahendrakar. Licensed under CC BY 4.0.

the reversal curse is "not an inherent failure but an artifact of how models encode knowledge", and the success of editing depends on representational geometry and not only on the editing algorithm.

The second conclusion arrives at a moment when the editing literature has been accumulating evidence that edits do not propagate. Edited models recall the edit in the direction it was written and fail in the reverse direction (Xu et al., 2023), fail on multi-hop questions that the edit entails (Zhong et al., 2023), and fail to make the inferences an injected fact licenses unless the inference shares words with the fact (Onoe et al., 2023). If propagation is a property of geometry, the field has been tuning the wrong component.

This paper takes the anchor study seriously and asks what it establishes. The answer it reaches has three parts.

First, the anchor's test is a test of relation-level inverse inference: given that B's son is C, infer that C's father is B, when the pairing of son with father has been shown on 500 complete families. That is the central inference of knowledge-graph completion, and the knowledge-graph community found it so easy that it had to be removed from its benchmarks: a model that does nothing but read inverse pairs from the training graph reached state-of-the-art scores on WN18 and FB15k (Dettmers et al., 2018; Akrami et al., 2020). The reversal curse as Berglund et al. (2023) posed it is a different problem. Its reverse query starts from a description that is not an entity with any other role in training, and the auxiliary data that showed other facts in both orders did not help.

Second, relation-level inverse inference is nevertheless not settled for next-token transformers, and this is where the anchor contributes. Two other from-scratch studies with the same design logic report near-total failure. On a dataset of 1,000 entities and 20 relations with their reverses, in which 39,800 training sequences demonstrate the relational pattern and 200 held-out facts are trained in one direction only, standard next-token training gives reverse accuracy near zero (Coda-Forno et al., 2026). With entities written as two-token names, standard transformers reach 0.0 percent on held-out reversals regardless of depth or embedding tying, although the same architecture reaches a mean reciprocal rank of 0.81 to 0.975 when each entity has its own embedding (Wang & Sun, 2025). The latter study trained with a weight decay of 0.25, which lies in the range where all of the anchor's models fail. The disagreement among these results is the live question, and the published differences between them are several, of which regularisation is only the one the anchor varies.

Third, "bilinear" names two different things in this literature. Zhu et al. (2024) prove the reversal curse in a bilinear model in which each token has one fixed embedding in both positions; the curse there comes from the learned matrix, which training raises only in the observed direction, even though reversal is demonstrated on every other pair. The anchor finds that bilinear structure lifts it. A shared encoding therefore does not by itself separate the two results. What the paper derives is a property of the anchor's probe: a bilinear score delivers an inverse relation by transposition exactly when the subject and the object are read through the same encoding. Knowledge-graph embedding found that condition missing in canonical tensor decomposition and repaired it by tying embeddings (Kazemi & Poole, 2018), and the anchor's probe imposes it by construction, since it extracts both entity vectors with the same function. The probe therefore shows that the hidden states admit a role-invariant bilinear reading. It does not show that the model's own computation uses one.

The paper makes six contributions. It separates fact-level reversal, relation-level inverse inference and edit propagation as three questions with different evidence bases, and reads the anchor's design against each (Sections 2 to 4). It compares the designs of four from-scratch and fine-tuning studies of reversal and identifies the variables on which the positive and negative results differ (Section 5, Table 2). It derives the role-sharing condition under which a bilinear score supports inversion by transposition, a condition the anchor's probe imposes and the bilinear model that predicted the curse already meets, so that it does not by itself explain the anchor's success (Section 6). It analyses what a post-hoc relational probe and a cross-model correlation can certify about mechanism (Section 7). It reads the anchor's editing result against the propagation literature in pretrained models (Section 8). Finally, it consolidates the published values the argument rests on in a table and a figure (Section 9), states a reading procedure as an algorithm (Section 10), and names six measurements that would settle the open part (Section 13).

2. Terms

Fact-level reversal. A model is trained on a statement linking A to B in one order and queried in the other order, where no other training statement links A to B and at least one of the two terms has no other role in training. This is the setting of Berglund et al. (2023) and of the inverse search task of Allen-Zhu and Li (2023b). Success requires information about the specific pair that the trained direction did not write in a form the reverse query can read.

Relation-level inverse inference. A model is trained on a statement (B, r, C) and queried for (C, r', B), where r' is the inverse of r and the pairing of r with r' has been demonstrated on many other entity pairs. In knowledge-graph completion this is the inverse or reciprocal pattern (Sun et al., 2019; Lacroix et al., 2018). Success requires a rule over relations and an entity representation the rule can act on.

Role-shared encoding. An entity has a role-shared encoding if the vector the model uses for it is the same, up to a fixed map, whether the entity occupies the subject slot or the object slot of a fact. In a next-token model the subject is read through input embeddings and the object is produced through output logits; these are

different roles.

Bilinear score. A function s(x, r, y) = f(x)ˆT M_r g(y), where f and g encode entities and M_r is a relation matrix. RESCAL takes f = g (Nickel et al., 2015). DistMult restricts M_r to be diagonal and is therefore symmetric in x and y (Yang et al., 2014). ComplEx moves to complex vectors so that antisymmetric relations can be represented (Trouillon et al., 2016).

Probe and mechanism. A probe is a model fitted after the fact to read a property from hidden states. A mechanism is the computation the model performs. A probe that succeeds shows that the property is decodable from the states it reads; whether the model uses that property is a separate question (Belinkov, 2022; Ravichander et al., 2021).

Propagation. After an edit to one fact, the change in the model's answers to queries whose correct answer is logically entailed by the edited fact. The RippleEdits benchmark calls the inverse and symmetric cases logical generalisation and the multi-hop cases compositionality (Cohen et al., 2023).

Memorising and generalising solutions. A network that fits its training set can do so with a solution that stores each example or with one that implements the rule that generated them. In grokking the second appears long after the first, and its appearance depends on dataset size and on weight decay (Power et al., 2022; Liu et al., 2022; Varma et al., 2023).

3. The anchor study, read at the resolution of its design

3.1 Data

The dataset has 1,000 families of 10 members each, related by eight relations: husband, wife, father, mother, son, daughter, brother and sister (Kim et al., 2025). Every family yields 36 facts. Each member has a name of the form first name, middle name, last name, and all members of a family share the middle and last name, which together act as a family name. Each fact is a short sentence of the form "[subject first name] [family name] [relation] [object first name] [family name]".

The families are split into two groups of 500. Group 1 contributes all 36 facts per family. Group 2 contributes 24, with father and mother removed. The test set consists of the removed facts, 12 per Group 2 family. A family is one training document in which all its facts are concatenated, and the training data contains 1,000 random orderings of each family's sentences per epoch, about 318 million tokens. Models are trained for 20 epochs. On the author's arithmetic from those two figures, each training fact is presented about 20,000 times.

3.2 What a test query requires

Consider a Group 2 family in which Jacob has a son Dalton. The training data contains "Jacob ... son Dalton ..." and the test asks "Dalton ... father". The answer is the family member X such that X's son is Dalton and X is male. Within the family, both facts needed are present: the son relation from Jacob to Dalton, and Jacob's gender, which is recoverable from the first-name pools (drawn separately by gender) and from the husband and wife relations. A second route also exists: Dalton's mother can be identified as the female parent through her own son fact, and her husband is Dalton's father. All of this is available inside a single family, and the target's family name is already in the prompt.

So the test asks for a relation that never appeared for Group 2 entities, which is genuinely out of distribution for those entities, but it does not ask for information about a pair that the training data did not contain. The pair (Jacob, Dalton) is present, in the son direction. What is missing is the father direction for that pair, together with the rule, demonstrated on 5,000 Group 1 entities, that son in one direction implies father or mother in the other.

3.3 Training outcomes

Weight decay is swept over nine values from 0 to 6.0, with three seeds each, giving 27 models trained with AdamW at a learning rate of 3×10^{-4}. All 27 reach 100 percent training accuracy. Test accuracy depends on weight decay and seed: below 1.0 the models consistently fail, and above it some seeds succeed and some do not. For analysis the authors select a representative "reversal cursed" model with reverse accuracy below 40 percent and a "not reversal cursed" model above 98 percent.

3.4 Probes

For each fact, the authors take the hidden state at the last token of the subject's name and at the last token of the object's name in a given layer, and fit three probes: a linear relational embedding in the sense of Hernandez et al. (2024), a translation vector in the sense of Merullo et al. (2024), and a ridge-regression variant of RESCAL that learns one matrix per relation. Probes are fitted on facts from 125 Group 1 families and evaluated on 125 other Group 1 families. A probe that ignored the subject but recognised the family and the relation could narrow each answer to three candidates in the family, so the stated chance level is one third.

In the successful model the bilinear probe rises sharply in layers 6 to 9 and peaks above 95 percent, while the linear and translational probes stay near or below chance; translational structure appears only for the symmetric husband and wife pair. In the failing model all three probes stay low. Some failing models show weak linear relational maps for individual relations, inconsistent across relations and layers. Algebraic tests on the fitted matrices show, in the successful model, that the transpose of a relation's matrix scores its inverse and that the product of two matrices scores their composition, with peaks in the same layers.

3.5 Editing

For each model the authors edit 50 husband facts, (A, husband, B) to (A, husband, B'), where B' is another member of the same family. Each edit fine-tunes only the output weights of the feed-forward block in one layer, with Adam at a learning rate of 4×10^{-4}, stopping when the loss falls below 0.2 or after 50 steps; this is repeated for each of the 12 layers. Three metrics are reported: success on the edited fact; logical generalisation, the mean success on (B', wife, A), on the father relation from A's children to B', and on the son or daughter relation from B' to those children; and locality, measured on other facts in the edited family that do not involve B' and on facts from untouched families.

Both models reach near-perfect edit success. The successful model generalises the edit and the failing model almost entirely does not, and the successful model also preserves unrelated facts better. Across models, peak bilinear probe accuracy and peak logical generalisation are related with R squared of 0.939. The best layers for editing are 1 to 4, while the bilinear structure is strongest in layers 6 to 9.

3.6 What the study does not report

The paper does not say in its text whether test queries are posed alone or after other facts from the same family, but its appendix presents the test set as "12 prompts per family", each a single fact sentence, which suggests that queries are posed alone. Because every training document concatenates a family's facts in random order, in Group 1 the father fact and its son counterpart often occur in the same context, so a model could learn to read a converse statement from earlier in the document. If test queries are posed alone, as the appendix suggests, that route is closed and the result concerns knowledge in the weights, which is how the authors interpret it. The paper does not report results for a training set in which facts appear in separate documents. It does not report the B' entity's prior relations, although B' is drawn from the same family and therefore already stands in some relation to A and A's children, and the edit makes the graph inconsistent unless those relations are revised too. Finally, its authors state that whether such structures exist in large pretrained models remains open.

4. Two questions called the reversal curse

4.1 The original design

Berglund et al. (2023) built three subsets of 30 fictitious celebrities each, every fact paraphrased 30 times into 900 documents per subset. One subset puts the name first, one puts the description first, and a third, used only for training, states each fact in both orders in separate documents so that the model might learn the pattern that facts appear both ways. Facts about different celebrities never share a document. GPT-3 models fine-tuned through an API after a hyperparameter sweep answer in the trained direction and are close to zero in the reverse one. The authors list what they tried to make the reverse generalise, including the both-orders subset and multiple paraphrases, and report that none of it helped.

Two features of this design matter for what follows. The description in "the first woman to travel to space" is not an entity with other facts; it occurs only as the other half of this one pair. And the both-orders subset demonstrates a pattern about the copula "is", not about a family of distinct inverse relations each with its own name.

Allen-Zhu and Li (2023b) obtain the same result in pretraining from scratch on synthetic biographies. Their inverse search task asks which person has a given birth date and city; the attribute values are not entities with their own facts. Inverse search stays at virtually zero regardless of training method, data or model size unless the biographies themselves state the knowledge in reverse order.

4.2 The knowledge-graph design

A knowledge graph treats both ends of a fact as entities, each with its own representation, and relation types as a small vocabulary. Inference of (C, r', B) from (B, r, C) is the inverse pattern, and models that can express it, such as RotatE's rotations or ComplEx's conjugate products, infer it routinely (Sun et al., 2019; Trouillon et al., 2016).

The community discovered that this pattern was too easy rather than too hard. Toutanova and Chen (2015) observed that FB15k contained many inverse pairs and built FB15k-237 without them. Toutanova and Chen (2015) also measured the problem: 94.0 percent of WN18 test triples and 80.9 percent of FB15k test triples have an entity pair that is "directly linked in the training knowledge graph" by some relation, figures that Dettmers et al. (2018) report as test leakage through inverse relations. Dettmers et al. showed that a rule-based model that does nothing but exploit inverse relations achieved state-of-the-art results on many metrics for both datasets, which led them to release WN18RR. Akrami et al. (2020) report that 70 percent of FB15k training triples form reverse pairs, that 70 percent of its test triples have a reverse in training, that the figures for WN18 are 92.5 and 93 percent, and that a simple rule of the form "(h, r1, t) implies (t, r2, h)" reaches filtered Hits@1 of 71.6 percent on FB15k and 96.4 percent on WN18. Lacroix et al. (2018) went further and made their tensor model invariant to whether a relation or its reciprocal is included in the data at all.

The anchor's test belongs to this class. Every Group 2 test fact has its inverse counterpart in training, and the inverse pairing is demonstrated on all of Group 1. On a knowledge-graph benchmark this would be called leakage. That does not make the experiment uninformative, because a next-token transformer is not a knowledge-graph embedding model, and whether it can learn the rule at all is a real question. Theory predicts that it may not: Zhu et al. (2024) prove the curse for a bilinear next-token model in exactly this design, in

which every other pair is trained in both orders, when the token embeddings are fixed and near-orthonormal (Section 6.1). It does mean that success on this test is evidence about relation-level inverse inference, and that extending it to fact-level reversal requires an argument the study does not supply.

4.3 Where the two questions meet

The two questions are not unrelated. Lin et al. (2024) find that models fine-tuned on biographies can answer reverse questions posed as multiple choice, where both terms appear in the prompt, but only when the training documents put the name first, and they trace this to a bias in how models begin recall from names in the question. Converse relations posed as multiple-choice matching against knowledge-graph triples remain hard for several model families, which often fall back on shortcuts (Qi et al., 2023). Bidirectional encoders trained on the same facts do not show the curse in a simple setting (Wu et al., 2023). Training on input-output pairs of a function can let a fine-tuned model state the function and compute its inverse (Treutlein et al., 2024), which is relation-level inversion of a rule, not recovery of an arbitrary pair. A fine-tuned model can generalise to implications of a new fact through structures learned in pretraining, but only if facts preceded their implications in that pretraining (Feng et al., 2024).

The common thread is that reverse success appears when the reverse query can reach an entity the model already represents and a rule it has already learned. That is the anchor's setting by design.

5. Where the from-scratch results disagree

The anchor's result is not in conflict with Berglund et al. (2023), whose task it does not test. It is in conflict with studies that test what it tests. Table 2 sets the designs side by side.

Six variables separate the positive result from the negative ones.

Regularisation. The anchor varies only this, and it matters within the anchor: no model with weight decay below 1.0 succeeds. Wang and Sun (2025) trained at 0.25. Coda-Forno et al. (2026) take their architecture and training setup from Lampinen et al. (2025b) and do not state a weight decay. On the published evidence it is therefore possible that the surface-level negative results were obtained in the regime where the anchor's own models also fail. No published study runs the negative designs at weight decays of 1.0 or more, so this is a candidate explanation, not a finding. Nor need the dependence be a simple threshold. In fine-tuning a pretrained 1-billion-parameter model on identity-bridge data, Ma et al. (2026) find a band of weight decay that helps reversal, with mean reverse accuracy of 0.12 at 0.15, a peak of 0.55 at 0.30 and 0.33 at 0.45, so that larger values hurt again. If the same holds from scratch, the candidate explanation rests on a non-monotonic dependence on regularisation, not on a single cut-off.

Entity encoding. Wang and Sun's contrast is the cleanest in the literature: the same transformer that reaches a mean reciprocal rank above 0.8 when entities are atomic embeddings reaches 0.0 percent when they are two-token names that share tokens with other names. They conjecture, and support with targeted experiments, that the difference comes from two failures of conceptual binding: the same entity receives inconsistent representations in different contexts, and representations of different entities become entangled through shared surface tokens. The anchor's names share far more surface than Wang and Sun's: every member of a family shares two of the three name parts, so the family name carries no information about which member is meant. The anchor nonetheless succeeds, which suggests either that strong regularisation overcomes the binding problem or that the closed family structure supplies the binding in another way.

The candidate set. Every test answer in the anchor is a member of the subject's own family, and the family name is in the prompt. The model must choose a first name from the ten members of one family, and for father only from its male members. Wang and Sun's and Coda-Forno's test answers are drawn from thousands of entities.

Document structure. The anchor concatenates a family's 36 facts into one document and reorders them 1,000 times. Converse facts therefore co-occur in context throughout Group 1 training. Zhu et al. (2024) note that when two facts are concatenated in one training sequence a model may learn the link directly, which their separate-sequence analysis excludes. Allen-Zhu and Li (2023a) find that knowledge is reliably extractable from a model only if it was augmented during pretraining, and they list sentence shuffling among the augmentations that work; the anchor's 1,000 orderings per family are such an augmentation. Berglund et al. (2023) keep facts about different people in separate documents, and in the simple reversal dataset used by Coda-Forno et al. (2026) each training sequence is a single relation sentence. The dataset's originating paper, however, also includes "12.5% of ICL sequences that have both forward and backward relations in the context" (Lampinen et al., 2025b), and Coda-Forno et al.'s appendix does not say whether they kept them. If they did, converse facts co-occurred in context in that negative design as well, which would weaken document structure as an explanation.

Alternative paths. Section 3.2 showed that each test fact can be reached by at least two routes inside a family. Lampinen et al. (2025a) report above-chance reversal from fine-tuning on a semantic structure, unlike on isolated facts, and suggest that other statements about an entity provide an alternative route. A dense, closed family graph is the extreme case of such structure.

Duration. At about 20,000 presentations per fact, the anchor trains far past the point of fitting the training set, which all its models reach. Wang et al. (2024) find that transformers learn implicit composition and comparison

Table 2. Designs of studies that train on one direction of held-out facts while showing the relational pattern on other pairs. "Pattern" is how the inverse is demonstrated elsewhere in training; values are as reported by each source.

Study	Entities	Pattern shown	Same document?	Regularisation	Reverse result
Kim 2025	3-part names, family name shared	8 relations, 500 full families	Whole family	AdamW, wd 0-6.0	Fail below wd 1.0; some seeds above 98%
Wang 2025	1 embedding per entity	6 relation pairs, 2.5K-100K entities	One fact	AdamW, wd 0.25	MRR 0.810-0.975
Wang 2025	2-token names, 10 share a token	Same, 50K entities	One fact	AdamW, wd 0.25	0.0%
Coda-Forno 2026	1,000 entities	20 relations plus reverses, 39,800 sequences	One fact	Not stated	Near 0% (NTP)
Berglund 2023	Names and descriptions	30 facts in both orders	No	API sweep	Near 0%

over parametric knowledge only through grokking, extended training long after overfitting. The anchor's dependence on weight decay and its split outcomes across seeds are the signatures of that regime (Section 7.3).

Each of these is a difference between designs, and few of them have been varied with the others held fixed. The anchor sweeps weight decay. Wang and Sun (2025) hold the transformer and task fixed and vary entity encoding, atomic embeddings against surface names, and also name overlap, reporting that "increasing the multiplicity severely hurts generalization". Ma et al. (2026) sweep weight decay, as above, and entity token length: with one-token number names reverse accuracy reaches nearly 100 percent, and with three-token names it falls to 10 percent. Two other recent results confirm that changing the training data or the objective can lift relation-level reversal in next-token models without a geometric claim. Fine-tuning a pretrained 1-billion-parameter model (Llama-3.2-1B-Instruct, on 100 husband-wife or parent-child pairs) with identity statements rewritten as a round trip through the forward and inverse relation, such as "The wife of Alice's husband is Alice", lifted it from near-zero reversal to about 50 percent, while the plain form "The name of Alice is Alice" gave no reversal; in the authors' words, "the identity bridge does not work without reforming to its OCR form" (Ma et al., 2026). The same study proves that under this data recipe even a single transformer layer can learn the reverse direction (Ma et al., 2026). Making the source entity a prediction target, by masking it in the training sequence, raises reversal accuracy on the simple reversal dataset to 99.5 percent for a decoder trained with masking and to 100 percent for a masked language model; forbidding the source to be masked returns it to 0 percent on all three datasets tested (Coda-Forno et al., 2026; Figure 1a). The anchor's models never see the father direction for Group 2 pairs and so never predict the source of the test query for those pairs. If their success is real in the in-weights sense, it is a relation-level rule acting on entity representations, which the source-prediction result neither predicts nor excludes, because Coda-Forno et al. test necessity within a design in which next-token training had already failed.

6. Which bilinear?

6.1 A bilinear model that has the curse

Zhu et al. (2024) analyse a bilinear model in which the logit for token B given token A is v_Aˆ T Theta v_B, where each token has one fixed embedding v, used whether it is the input or the predicted token, and Theta is a learned matrix. The training set contains n pairs; for all but one of them both orders are trained, and the remaining pair is trained in one order and tested in the other, so reversal is demonstrated on every other pair. When embeddings are near-orthonormal and training sees A followed by B, gradient descent raises the entry of Theta that links A to B and leaves the entry linking B to A essentially untouched. The authors name this asymmetry of the learned matrix as a core reason for the curse: the logit for A given B does not rise, and the model has the reversal curse. They show the same mechanism in one-layer transformers and add a second property, intransitivity: training A to B and B to C does not raise A to C.

6.2 A bilinear structure that lifts it

The anchor's bilinear probe scores a fact as h_sˆT W_r h_o, where h_s and h_o are hidden states taken in the same layer, at the last token of each name, by the same extraction procedure. The algebraic test asks whether W_rˆT scores the inverse relation.

6.3 The condition the probe needs

Write a general bilinear score as s(x, r, y) = f(x)ˆT M_r g(y). Its transpose gives f(y)ˆT M_rˆT g(x). Because both expressions are scalars, the second equals g(x)ˆT M_r f(y). For the transposed matrix to score the inverse relation on every pair, as the inverse of a relation must, this has to equal the forward score with the arguments swapped, which requires g(x)ˆT M_r f(y) = f(x)ˆT M_r g(y) for all x and y. The condition holds when f = g, the role-shared encoding of Section 2, and in general fails when f and g are independent.

This is a derivation, not a measurement, and it says three things.

First, it says what the condition does not explain. In Zhu et al.'s model each token has one fixed embedding in both positions, so f = g and the condition holds, yet the

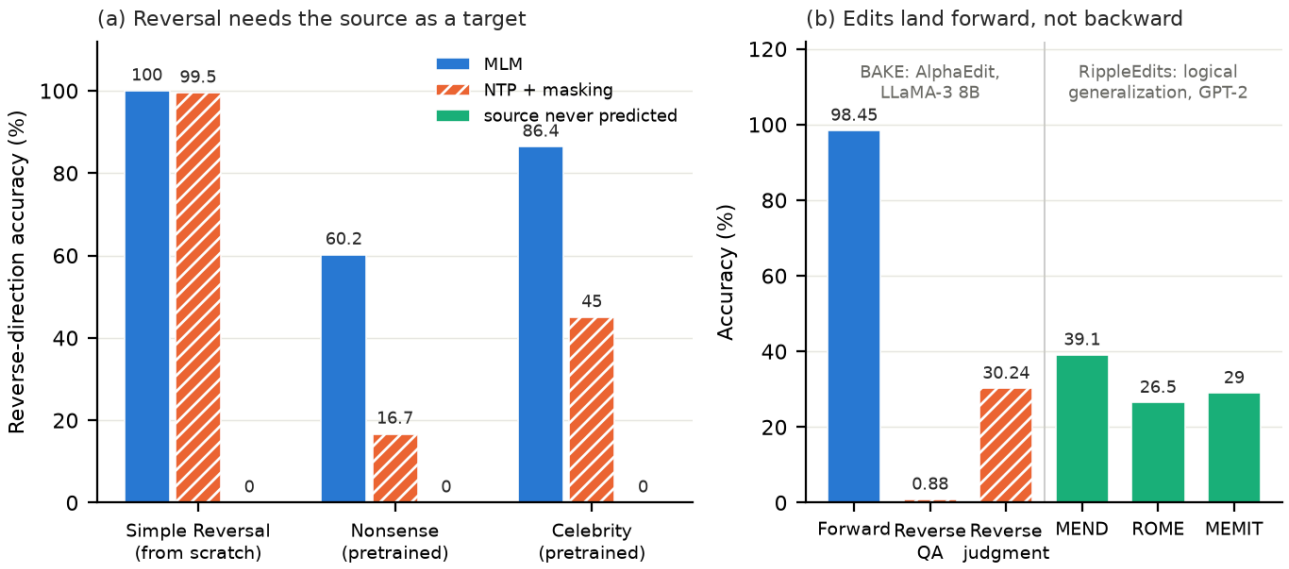


Figure 1. (a) Reverse-direction accuracy under three training objectives on three reversal datasets, from Coda-Forno et al. (2026), Table 1: a masked language model, a decoder trained with masking, and an ablation in which the source entity is never masked. (b) Accuracy of edited pretrained models: AlphaEdit on LLaMA-3 8B in the BAKE benchmark, forward efficacy and reverse-direction question-answering and judgment scores (Xu et al., 2023, v4); logical-generalisation accuracy of MEND, ROME and MEMIT on GPT-2 in RippleEdits, averaged over three splits (Cohen et al., 2023). All values are as published by the named sources.

model has the curse: the reverse logit is read through the transpose of Theta, but training on the held-out pair never raises the reverse entry. A shared encoding makes transposition read as inversion; it does not make a learner write the inverse into the transposed matrix. The condition is therefore a property of the anchor's probe, not what separates Zhu et al.'s result from the anchor's. In the anchor's probe, f and g are the same extraction function, so transposition reads the same encodings in the other order, and the transpose test asks whether the fitted matrices carry the inverse.

Second, the knowledge-graph literature met exactly this condition. Canonical polyadic decomposition, a bilinear factorisation, performed poorly for link prediction because "it learns two independent embedding vectors for each entity, whereas they are really tied"; SimplE repaired it by letting the two embeddings of each entity be learned dependently (Kazemi & Poole, 2018). RESCAL, which the anchor's probe adapts, uses one embedding per entity in both roles (Nickel et al., 2015). DistMult showed that composition of relations in such models is characterised by matrix multiplication (Yang et al., 2014), which is the algebraic property the anchor's composition test checks.

Third, it connects the anchor's finding to Wang and Sun's (2025) diagnosis. Their first conjectured cause of the curse is that transformers lack an inductive bias toward consistent representations of a concept across the contexts in which it appears. A role-shared encoding is a consistent representation across the two roles that matter for reversal. Their concept-level experiment, in which each entity is a single learnable embedding and predictions are scored over the embeddings of all concepts, comes close to satisfying the condition by design and succeeds; their surface-level experiment does not and fails. Diffusion language models, trained bidirectionally, still show the curse, and an analysis attributes part of it to entities fragmented across separately predicted tokens (He et al., 2026), another failure of consistent encoding. A theoretical account of why masked diffusion models suffer less from the curse traces the gain to evidence stored in parameters that the forward and reverse positional conditionals both use (Jeon et al., 2026).

6.4 What the probe presupposes

The anchor's probe applies one extraction function to both ends of a fact and fits a matrix. If it succeeds, the hidden states at those positions admit a role-shared bilinear reading. It does not follow that the network computes the father relation by applying the transpose of a son matrix to a role-shared code. The network's own output is produced by attention and feed-forward computation over a sequence and a vocabulary readout in which the object is a predicted token, not a vector extracted at a later position.

One further detail bears on what the probe's role-sharing establishes. The entity vectors are described as taken for each fact at the last token of each name, but the fitting objective (Kim et al., 2025, Eq. 5) builds its entity matrix from one subject-position state per entity, over the n = 1,250 training entities, and uses that matrix on both sides of the relation matrix. The fitted matrices are therefore estimated from per-entity subject states. That makes the probe role-shared by construction, and in fitting it does not read an object state conditioned on the pair being scored. The evaluation step is described less explicitly. If at evaluation the object's vector is taken

inside the fact sentence, it is computed after the model has attended to the subject and the relation, and is then conditioned on the very pair the probe is scoring. The failing models show that this alone would not make the probe succeed, since their states are computed the same way and the probe fails on them. Extracting each entity's vector from a neutral context at evaluation as well would test the stronger claim.

6.5 What would contradict the role-sharing reading

A relational geometry need not be role-shared to support reversal. Coda-Forno et al. (2026) find, for the objectives that succeed in their setting, that forward and reverse directions behave as distinct entries in representation distance and linear probes, with no single direction-agnostic representation. If the anchor's successful models also store the two directions separately, and the bilinear probe succeeds only because its own fitted matrices impose the transpose relation that the network implements in another way, then the role-sharing condition would describe the probe rather than the model. Chang et al. (2026) prove that guaranteeing first-order propagation under negation forces a tensor factorisation that separates entity-pair context from relation content, and that requiring it under the converse operation as well further forces the symmetric and antisymmetric components to align with argument order. That is a necessary condition on the geometry of updates, not a description of any trained model, and whether the anchor's successful models satisfy it has not been measured.

7. What the probe and the correlation can carry

7.1 Decodable is not used

The interpretability literature has spent several years establishing that a probe's accuracy is not evidence that the probed property drives behaviour. Probes can learn the task themselves, which control tasks are designed to expose (Hewitt & Liang, 2019). Models encode properties they do not need for their task, and properties distributed as random noise in the data can be decoded well above chance (Ravichander et al., 2021). The behavioural question is answered by removing the property and observing the output, as amnesic probing does (Elazar et al., 2021), not by decoding it. Belinkov (2022) surveys these limits.

The anchor's bilinear probe is a fitted model with one matrix per relation in a hidden space of 896 dimensions, trained on 1,250 entities. It is evaluated on held-out families, which guards against memorising entities, and compared with two other probes and with a chance level, which guards against some trivial readings. It is not compared with a control task, and no intervention on the probed structure is reported.

7.2 Layer mismatch

The anchor's editing results add a second reason for caution. Edits generalise best when made in layers 1 to 4, while the bilinear probe peaks in layers 6 to 9. The authors read this as editing the layers where the structured representation is formed rather than where it is used. That reading is consistent with a well-supported picture of factual recall in pretrained transformers, in which early feed-forward layers enrich the representation at the subject's last token and later attention heads extract the attribute (Geva et al., 2023), and with evidence that the best layer to edit is a better predictor of editing success than where causal tracing localises a fact (Hase et al., 2023). It is also consistent with a simpler reading: a fine-tuning step in an early layer changes the encoding of the entities in the edited sentence, and any downstream computation that uses those encodings, bilinear or not, then produces the entailed answers. The edit tests whether the downstream computation generalises over entity encodings. It does not test whether that computation is the bilinear form the probe found.

7.3 One solution, three symptoms

The R squared of 0.939 between peak probe accuracy and peak logical generalisation is computed across the anchor's models, which differ in weight decay and seed and which split into models that generalise and models that do not. The grokking literature describes what such a split looks like. Networks first fit the training set with a memorising solution and later, under weight decay, switch to a generalising one that is slower to learn but more efficient, giving larger logits for the same parameter norm (Varma et al., 2023; Liu et al., 2022). The transition is gradual underneath and can be tracked by mechanistic progress measures (Nanda et al., 2023). Whether it happens depends on dataset size and regularisation, and it can be partial or reversed (Varma et al., 2023). In synthetic studies of composition and comparison over parametric knowledge, transformers learned implicit reasoning only through this route (Wang et al., 2024).

A model that has switched to the generalising solution will infer the withheld relations, will carry structure that a relational probe can read, and will propagate an edit to an entity encoding through the relational computation. A model that has not will do none of the three. A tight correlation across models between any two of these symptoms is what the presence or absence of one solution predicts, whatever its internal form. The correlation therefore supports the claim that propagation and inverse inference come from the same learned structure. It does not support the claim that the structure is bilinear rather than some other relational computation, because a model with a different generalising mechanism would also, by hypothesis, carry some readable structure.

8. Propagation in a closed world and in pretrained models

8.1 What the anchor's editing scores

The anchor's edit changes A's husband from B to B', and logical generalisation is scored on facts that should now hold: B' has wife A, A's children have father B', and B' has those children. Three kinds of entailed change are not scored. Facts that should no longer hold, such as B having wife A, are not scored, except where the father relation is functional, so that answering B' implies not answering B. Facts that conflict with B"s existing relations are not scored either: B' was drawn from the same family, so B' already stands in some relation to A and to A's children, as a son or a brother for instance, and the edit makes the graph inconsistent unless those relations change. The in-family locality metric excludes facts involving B', so those conflicts are neither counted as failures nor as preserved knowledge. And facts about B's other relations, which the edit may disturb, are counted only through locality.

This is a choice of which consequences define success, and Hase et al. (2024) argue that it cannot be avoided: model editing is belief revision, which entailed changes an edit should produce is underdetermined, and benchmarks that fix them encode a choice. Li et al. (2023) report that edits to groups of facts that conflict logically magnify existing inconsistencies in pretrained models, and that edits can distort the knowledge structure itself. Nishi et al. (2025) train transformers from scratch on a structured knowledge graph in which an edit has consequences for other entities, and find that editing distorts the representations of entities beyond the target, degrading recall and reasoning, with corroborating evidence in pretrained Llama and Mamba models. The anchor reports that its successful models preserve unrelated facts better than its failing ones, which is the opposite of shattering for its single fine-tuned edits; with batch or locate-and-edit methods the question is open.

8.2 What pretrained models show

The evidence on propagation in pretrained models is consistent and negative. Figure 1b shows two benchmarks. In BAKE, AlphaEdit on LLaMA-3 8B recalls 98.45 percent of edits in the edited direction, 0.88 percent in the reverse direction posed as a question, and 30.24 percent in the reverse direction posed as a judgment; the in-context editing baseline mitigates the gap to some extent (Xu et al., 2023). In RippleEdits, logical generalisation accuracy on GPT-2 is 39.1 percent for MEND, 26.5 percent for ROME and 29.0 percent for MEMIT, and an in-context editing baseline obtains the best scores (Cohen et al., 2023). On multi-hop questions whose answers an edit should change, weight-edited models fail badly, and storing the edits outside the model and prompting it step by step works better (Zhong et al., 2023). Injected facts propagate to inferences only where the inference shares words with the fact, while placing the definition in context helps across settings (Onoe et al., 2023), and context distillation propagates better than fine-tuning or gradient-based editing (Padmanabhan et al., 2023). Having a model write out the implications of seed documents and then fine-tuning on the implications it judges true raises accuracy by between 3 and 26 percent on three datasets, one of them the reversal-curse data (Akyurek et al., 2024). Edited models also overweight the edit target in a way that harms multi-hop use (Zhang et al., 2024), editing relations rather than entities is harder and touches attention as well as feed-forward layers (Wei et al., 2024), and in lifelong editing an interference term arising from superposition of knowledge representations explains why edits damage other knowledge (Hu et al., 2024).

What does predict propagation in pretrained models is a geometric quantity, but a different one from the anchor's. The cosine similarity between the gradient of an edited fact and the gradient of a related fact correlates strongly and positively with whether the edit ripples to that fact, across models, editors and metrics, and the failure cases of negation, over-rippling and cross-lingual queries have low gradient similarity (Qin et al., 2024). A representation-level measure of entanglement between facts improves Spearman correlation with ripple effects by 62.2 percent on average over the gradient-similarity baseline (GradSim) across a corpus of 11,427 facts (Baser et al., 2026). Chang et al. (2026) measure the gradient of a fact against the gradient of its negation in Qwen3-4B, Qwen3-30B and OLMo-3-7B and find them strongly positively aligned, where systematic propagation would require them to point in opposite directions.

These results support the anchor's general thesis that propagation is a property of geometry and not only of the editing algorithm. They do not bear on its specific mechanism. What they measure is tangent-space or activation-space similarity between facts, which governs which facts an update moves together; it is not a test for a relation matrix whose transpose is its inverse. Whether pretrained models contain such matrices for any relation is not known. The nearest evidence is that relation decoding in pretrained models is well approximated by a single affine map for some relations, robustly for 48 percent of the relations tested, chiefly on GPT-J (Hernandez et al., 2024), that some relational tasks reduce to an additive update in feed-forward layers (Merullo et al., 2024), that attribute extraction typically runs through attention heads that encode subject-attribute mappings in their parameters (Geva et al., 2023), that recall combines several independent additive mechanisms (Chughtai et al., 2024), and that the operations attention heads implement can be mapped from their parameters, for 20 predefined operations across six models (Elhelo & Geva, 2025). None of these tests a role-shared bilinear structure.

8.3 The theoretical limit on single updates

Chang et al. (2026) prove a stronger negative result for the first-order regime. If a linear update to the features of one fact is to propagate systematically to its negation and converse, the features must factorise into entity-pair context and relation content; and if composition is also to propagate, conjunction on those features must be bilinear, which is incompatible with the negation constraint and forces the features to collapse. The result is local: it governs single small updates and the lazy-training regime, and it need not apply to non-local optimisation. The anchor's edits are up to 50 fine-tuning steps on a family graph with no negation, so neither of the theorem's conditions binds there. The theorem does say that the anchor's success, if it transfers, will not transfer to every logical operation at once by a single local update.

9. Consolidated evidence

Table 1 collects the published values that the argument rests on, one row per result, with the source named on each row.

Three patterns stand out. First, among the from-scratch studies in Table 2, every success at reverse inference without reversed data comes from a design in which the inverse is demonstrated on other entities: Kim et al.'s generalising models, and Wang and Sun's atomic embeddings, which give each entity a single code by construction. Outside Table 2, knowledge-graph embedding models with one embedding per entity make the same inference so readily that it had to be removed from their benchmarks. The converse does not hold: demonstrating the inverse on other pairs is not enough, since Zhu et al.'s (2024) theory predicts failure in exactly that design when embeddings are fixed and near-orthonormal. Second, every external negative result in that design was obtained at a weight decay below 1.0 or at one the study does not report. Within the anchor, strong weight decay is necessary but not sufficient: above 1.0 some models succeed and others do not, and the outcome depends on the seed as well as on the weight decay. Third, every measurement of propagation in pretrained models collected here falls well short of complete propagation, and the best results come from placing information in context or from writing consequences into the training data, not from editing weights.

10. A procedure for reading a reversal or propagation claim

Algorithm 1 states the questions a reader can ask of any claim that a model infers reverse facts or propagates an edit, in the order in which they discriminate.

Applied to the anchor, the procedure returns relation-level inverse inference, with role sharing imposed by the probe and untested in the model; the target never predicted from the query subject for test pairs; family facts sharing documents; at least two forward paths per test fact; strong weight decay and about 20,000 presentations per fact; a candidate set of one family; and a mechanism status of decodable. Its edits score entailed additions, partly score retractions through functional relations, and leave conflicts with B"s prior relations unscored. Applied to Berglund et al. (2023), it returns fact-level reversal with no consistent code for the description, no shared documents, no forward path, and a candidate set of all names.

The procedure does not decide whether a claim is true. It decides which question a claim answers and which evidence would move it from decodable to used.

11. The strongest case against this reading

The case for the anchor's broader conclusion is stronger than the preceding sections might suggest, and it deserves a statement in its best form.

First, the anchor withholds two whole relations for half the entities, which is harder than the usual knowledge-graph protocol of withholding scattered triples. The model never sees a father or mother fact for any Group 2 entity, so it cannot rely on having learned which entities tend to be fathers. That is a fair point, and it is why the result is evidence of a relational rule rather than of memorised entity statistics. It does not change the classification: every withheld fact still has its inverse in training.

Second, the transpose and composition tests are not trivial. A probe that merely memorised pairs would not produce matrices whose transposes score inverse relations on held-out families. That is correct, and it is the strongest evidence in the paper that something with relational algebra is decodable from the generalising models' middle layers. The question Section 7 raises is whether the network uses it, not whether it is there.

Third, the negative from-scratch results may simply have been underpowered, run at too little regularisation or for too few steps, in which case the anchor's result is the correct one and the curse is indeed an artifact of training choices. This reading is consistent with all of Table 2 and is the most important open possibility. If it holds, the anchor will have shown that next-token transformers can learn relation-level inverse inference over surface-form names, which is a significant result. It will still not have shown that they can perform fact-level reversal, because no amount of regularisation creates a training signal for a pair the data never linked in any form; that is the case Berglund et al. (2023) and Allen-Zhu and Li (2023b) test, and on which reverse training (Golovneva et al., 2024), permutation of semantic units (Guo et al., 2024) and bidirectional objectives (Kitouni et al., 2024; Nie et al., 2025) all act by putting the reverse into the training signal.

Fourth, in pretrained models some reversal knowledge plainly exists. GPT-4 answers many real reverse questions about celebrities' parents, if less often than forward ones (Berglund et al., 2023), and real text states many facts in both orders. The anchor's argument could

Table 1. Published values bearing on reversal, relation-level inference and edit propagation. Values are as reported by the named source; "wd" is weight decay, "NTP" next-token prediction, "LG" logical generalisation.

Source	Setting	Condition	Readout	Reported value
Kim 2025	Family KG, from scratch	wd below 1.0	Withheld parent relations	All models fail
Kim 2025	Same	Representative pair	Same	Below 40% / above 98%
Kim 2025	Same, generalising model	Bilinear probe, layers 6-9	Probe accuracy (chance 1/3)	Above 95%
Kim 2025	Models in the sweep	Probe vs editing	R squared	0.939
Kim 2025	Same	Best editing layers	Layer index	1-4 (probe peak 6-9)
Berglund 2023	GPT-3, fine-tuned	30 facts x 30 paraphrases	Reverse accuracy	Close to 0%
Allen-Zhu 2023b	Biographies, from scratch	Forward-only data	Inverse search	Virtually 0%
Wang 2025	Atomic entity embeddings	2.5K-100K entities	Held-out reversal MRR	0.810-0.975
Wang 2025	2-token names, wd 0.25	50K entities	Held-out reversal	0.0%
Coda-Forno 2026	Simple Reversal	MLM / NTP + mask / no source	Reverse accuracy	100 / 99.5 / 0%
Coda-Forno 2026	Celebrity, pretrained	MLM / NTP + mask / no source	Reverse accuracy	86.4 / 45.0 / 0%
Coda-Forno 2026	Nonsense, pretrained	MLM / NTP + mask / no source	Reverse accuracy	60.2 / 16.7 / 0%
Ma 2026	1B pretrained, fine-tuned	Identity bridge, round-trip form / plain form	Reversal success	About 50% / near zero
Toutanova 2015 (via Dettmers 2018)	WN18 / FB15k	Test entity pair linked in training	Test triples linked	94.0% / 80.9%
Akrami 2020	FB15k / WN18	Inverse-rule model	Filtered Hits@1	71.6% / 96.4%
Akrami 2020	FB15k / WN18	Test triples	Reverse in training	70% / 93%
Yang 2014	Freebase	Bilinear vs TransE	Top-10 accuracy	73.2% vs 54.7%
Hernandez 2024	GPT-J relations	Affine relation map	Relations with robust map	48%
Xu 2023 (BAKE)	AlphaEdit, LLaMA-3 8B	Forward / reverse QA / judgment	Accuracy	98.45 / 0.88 / 30.24%
Cohen 2023	GPT-2, RippleEdits	MEND / ROME / MEMIT	LG accuracy	39.1 / 26.5 / 29.0%
Akyurek 2024	Three datasets	Deductive closure training	Accuracy gain	3-26%
Baser 2026	11,427 facts	Entanglement measure	Spearman gain vs GradSim	62.2% average
Grosse 2023	Up to 52B	Key phrases flipped	Influence	Decays to near zero
Wang 2024	Grokked transformer	Composition, OOD	Systematic generalisation	Fails (comparison succeeds)
Varma 2023	Grokking tasks	Varied dataset size	Generalisation	Ungrokking and semi-grokking observed

be that pretrained models fail at reversal where their encoding of an entity is not role-shared, and succeed where real data has made it so. That is a testable hypothesis, and it is one of the measurements proposed in Section 13.

12. What is not known

Whether any large pretrained language model encodes any relation with a role-shared bilinear structure is not known; the anchor's authors say so, and no study located for this paper has tested it.

Whether the negative from-scratch results on surface-form names reverse at weight decays of 1.0 or more is not known.

Whether the anchor's successful models answer test queries from their weights when each query is posed alone, rather than with family context, is not stated in the text, although the appendix's single-sentence test prompts suggest that the reported accuracy is for queries posed alone.

Whether removing the bilinear structure from a successful model removes its ability to infer the withheld relations is not known. No intervention on the structure has been reported.

Whether the anchor's editing result holds for locate-and-edit or batch editors, which change weights in closed form rather than by gradient steps, is not known.

How the anchor's edited models answer queries whose prior answers conflict with the edit, such as B"s existing relations to A and A's children, is not reported.

Whether relation-level inverse inference learned on a closed family graph transfers to open-world relations where the candidate set is every entity the model knows is not known.

Algorithm 1. Reading a claim that a model infers reversed or entailed facts

```
Input: a training set D, a test query q = (y, r', ?) with gold answer x,
       and a claim that the model answers q from its weights
step 1: If (x, r, y) is absent from D in every form:
       the claim concerns knowledge the data never held; reject or reframe.
step 2: If (x, r', y) or a paraphrase of it is in D, or q's answer is in the
   prompt's context: label FORWARD RETRIEVAL or IN-CONTEXT; stop.
step 3: If the pairing (r, r') is demonstrated on other pairs in D:
       label RELATION-LEVEL INVERSE INFERENCE
   else:
       label FACT-LEVEL REVERSAL
step 4: Record the conditions that the published record links to success:
   a. Is x encoded consistently as subject and as object (role sharing)?
   b. Is x ever a prediction target given y, or only given others?
   c. Do facts about x and y share training documents?
   d. Does a forward path reach x from y through other relations?
   e. What regularisation and training duration were used?
   f. How large is the candidate set for x at test time?
step 5: If a structure S is offered as the mechanism (bilinear, linear, etc.):
       require an intervention that removes or alters S and changes the
       answer to q; a probe or a cross-model correlation alone gives
       DECODABLE, not USED.
step 6: If an edit is claimed to propagate:
   a. List the entailed additions, retractions and conflicts.
   b. Report each separately; treat unscored categories as unknown.
   c. Compare with an in-context baseline holding the edited fact.
Output: the label from step 2 or 3, the conditions from step 4, and the
        evidential status of the mechanism (USED or DECODABLE).
```

13. Six measurements that would settle the open part

M1. Weight decay on the negative designs. Train the surface-name design of Wang and Sun (2025) and the simple reversal design of Coda-Forno et al. (2026) at the anchor's weight decays and durations. If reverse accuracy rises from zero, regularisation is the operative variable and the curse in this setting is a training artifact. If it stays at zero, something in the anchor's family structure is doing the work.

M2. The anchor without its family structure. Retrain the anchor's design with each fact as its own document, with answers drawn from all entities rather than from a family, and with family names removed so that names do not share parts. Each change removes one of the variables of Section 5. Reporting reverse accuracy after each change would locate the effect.

M3. Isolated test queries. Confirm that the anchor's reported test accuracy is for each query posed alone, as its appendix suggests, and report it as well with the family's other facts in context. A gap would show how much of the result is in-context converse reading learned from concatenated documents.

M4. Intervention on the probed structure. In a generalising model, project out the subspace the bilinear probe uses in layers 6 to 9, or replace it with the corresponding states from a failing model, and measure test accuracy. If accuracy falls, the structure is used; if it does not, it is decodable only. Amnesic probing provides the method (Elazar et al., 2021).

M5. A role-sharing test in pretrained models. For relations with a known inverse in a pretrained model, extract entity vectors from neutral contexts, fit a bilinear probe on forward facts, and test whether its transposed matrices score inverse facts on held-out entities. Then compare propagation of edits, measured as in RippleEdits or BAKE, between relations where the transpose test succeeds and relations where it fails. A difference would carry the anchor's thesis to pretrained models; no difference would confine it to the synthetic setting.

M6. Full consequence scoring of edits. Re-run the anchor's edits scoring three categories separately: entailed additions, entailed retractions, and conflicts with the new partner's prior relations. Add an in-context baseline in which the edited fact is placed in the prompt, as in RippleEdits and BAKE. This would show whether the generalising models revise beliefs coherently or only add the edited fact's direct consequences.

14. Limitations

This analysis rests on the published text of the anchor study and of the studies compared with it. The comparisons in Table 2 are between designs that differ in several ways at once, and the paper's attribution of the disagreement to six variables is an argument about which differences could matter, not a measurement of which do. The derivation in Section 6.3 concerns the transpose test as the anchor states it and does not exclude other ways in which a network could implement inverse relations. The reading of the anchor's probe in Section 6.4 relies on the extraction procedure as described in its text; if the authors extracted entity vectors in another way, that part of the argument would weaken. Several of the most relevant sources are 2025 and 2026 preprints that have not all been independently replicated. The search for this paper covered arXiv and the knowledge-graph completion literature and may have missed from-scratch reversal studies published elsewhere.

15. Conclusion

Kim et al. (2025) show that a next-token transformer trained from scratch can learn to infer a withheld inverse relation, and that when it does, its middle layers admit a bilinear relational reading whose transposes and products behave like inverses and compositions, and a fine-tuned edit propagates to entailed facts. Those are real results. They answer whether next-token training can learn relation-level inverse inference over entities with dense relational context, which is the inference knowledge-graph completion found easiest and the one other from-scratch studies with surface-form names have found impossible; the anchor's sweep over weight decay is the most specific clue yet to why the results differ. They do not answer the question the reversal curse posed, which concerns a single pair with no other link, and they do not show that the network computes with the bilinear structure that a role-sharing probe can read. The paper has stated the condition under which bilinear structure supports inversion, the comparisons that would assign the disagreement to its causes, and the intervention that would move the geometric claim from decodable to used.

References

Akrami, F., Saeef, M. S., Zhang, Q., Hu, W., & Li, C. (2020). Realistic Re-evaluation of Knowledge Graph Completion Methods: An Experimental Study. arXiv:2003.08001. doi:10.48550/arXiv.2003.08001.

Akyurek, A. F., Akyurek, E., Choshen, L., Wijaya, D., & Andreas, J. (2024). Deductive Closure Training of Language Models for Coherence, Accuracy, and Updatability. arXiv:2401.08574. doi:10.48550/arXiv.2401.08574.

Allen-Zhu, Z., & Li, Y. (2023a). Physics of Language Models: Part 3.1, Knowledge Storage and Extraction. arXiv:2309.14316. doi:10.48550/arXiv.2309.14316.

Allen-Zhu, Z., & Li, Y. (2023b). Physics of Language Models: Part 3.2, Knowledge Manipulation. arXiv:2309.14402. doi:10.48550/arXiv.2309.14402.

Baser, M., Yildiz, A., Divakaran, D. M., & Gurusamy, M. (2026). CLaRE-ty Amid Chaos: Quantifying Representational Entanglement to Predict Ripple Effects in LLM Editing. arXiv:2603.19297. doi:10.48550/arXiv.2603.19297.

Belinkov, Y. (2022). Probing Classifiers: Promises, Shortcomings, and Advances. Computational Linguistics. arXiv:2102.12452. doi:10.48550/arXiv.2102.12452.

Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A. C., Korbak, T., & Evans, O. (2023). The Reversal Curse: LLMs trained on 'A is B' fail to learn 'B is A'. arXiv:2309.12288. doi:10.48550/arXiv.2309.12288.

Chang, H., Mucsanyi, B., & Oh, S. J. (2026). Dynamics Reveals Structure: Challenging the Linear Propagation Assumption. arXiv:2601.21601. doi:10.48550/arXiv.2601.21601.

Chughtai, B., Cooney, A., & Nanda, N. (2024). Summing Up the Facts: Additive Mechanisms Behind Factual Recall in LLMs. arXiv:2402.07321. doi:10.48550/arXiv.2402.07321.

Coda-Forno, J., Wang, J. X., & Chaudhry, A. (2026). The Illusion of Latent Generalization: Bi-directionality and the Reversal Curse. arXiv:2604.04943. doi:10.48550/arXiv.2604.04943.

Cohen, R., Biran, E., Yoran, O., Globerson, A., & Geva, M. (2023). Evaluating the Ripple Effects of Knowledge Editing in Language Models. arXiv:2307.12976. doi:10.48550/arXiv.2307.12976.

Dettmers, T., Minervini, P., Stenetorp, P., & Riedel, S. (2018). Convolutional 2D Knowledge Graph Embeddings. AAAI 2018. arXiv:1707.01476. doi:10.48550/arXiv.1707.01476.

Elazar, Y., Ravfogel, S., Jacovi, A., & Goldberg, Y. (2021). Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals. arXiv:2006.00995. doi:10.48550/arXiv.2006.00995.

Elhelo, A., & Geva, M. (2025). Inferring Functionality of Attention Heads from their Parameters. ACL 2025. arXiv:2412.11965. doi:10.48550/arXiv.2412.11965.

Feng, J., Russell, S., & Steinhardt, J. (2024). Extractive Structures Learned in Pretraining Enable Generalization on Finetuned Facts. arXiv:2412.04614. doi:10.48550/arXiv.2412.04614.

Geva, M., Bastings, J., Filippova, K., & Globerson, A. (2023). Dissecting Recall of Factual Associations in Auto-Regressive Language Models. EMNLP 2023. arXiv:2304.14767. doi:10.48550/arXiv.2304.14767.

Golovneva, O., Allen-Zhu, Z., Weston, J., & Sukhbaatar, S. (2024). Reverse Training to Nurse the Reversal Curse. arXiv:2403.13799. doi:10.48550/arXiv.2403.13799.

Grosse, R., Bae, J., Anil, C., Elhage, N., Tamkin, A., Tajdini, A., et al. (2023). Studying Large Language Model Generalization with Influence Functions. arXiv:2308.03296. doi:10.48550/arXiv.2308.03296.

Guo, Q., Wang, R., Guo, J., Tan, X., Bian, J., & Yang, Y. (2024). Mitigating Reversal Curse in Large Language Models via Semantic-aware Permutation Training. arXiv:2403.00758. doi:10.48550/arXiv.2403.00758.

Hase, P., Bansal, M., Kim, B., & Ghandeharioun, A. (2023). Does Localization Inform Editing? Surprising Differences in Causality-Based Localization vs. Knowledge Editing in Language Models. NeurIPS 2023. arXiv:2301.04213. doi:10.48550/arXiv.2301.04213.

Hase, P., Hofweber, T., Zhou, X., Stengel-Eskin, E., & Bansal, M. (2024). Fundamental Problems With Model Editing: How Should Rational Belief Revision Work in LLMs? arXiv:2406.19354. doi:10.48550/arXiv.2406.19354.

He, S., Wei, K., Zeng, X., Chen, X., Yang, X., Li, Z., et al. (2026). DiffER: Diffusion Entity-Relation Modeling for Reversal Curse in Diffusion Large Language Models. arXiv:2601.07347. doi:10.48550/arXiv.2601.07347.

Hernandez, E., Sharma, A. S., Haklay, T., Meng, K., Wattenberg, M., Andreas, J., Belinkov, Y., & Bau, D. (2024). Linearity of Relation Decoding in Transformer Language Models. ICLR 2024. arXiv:2308.09124. doi:10.48550/arXiv.2308.09124.

Hewitt, J., & Liang, P. (2019). Designing and Interpreting Probes with Control Tasks. EMNLP 2019. arXiv:1909.03368. doi:10.48550/arXiv.1909.03368.

Hu, C., Cao, P., Chen, Y., Liu, K., & Zhao, J. (2024). Knowledge in Superposition: Unveiling the Failures of Lifelong Knowledge Editing for Large Language Models. arXiv:2408.07413. doi:10.48550/arXiv.2408.07413.

Jeon, M., Shin, S., Kim, B., Lee, K., & No, A. (2026). A Theoretical Analysis of Why Masked Diffusion Models Mitigate the Reversal Curse. arXiv:2602.02133. doi:10.48550/arXiv.2602.02133.

Kazemi, S. M., & Poole, D. (2018). SimplE Embedding for Link Prediction in Knowledge Graphs. NeurIPS 2018. arXiv:1802.04868. doi:10.48550/arXiv.1802.04868.

Kim, D.-K., Kim, M., Kwon, J., Yang, N., & Cha, M. (2025). Bilinear representation mitigates reversal curse and enables consistent model editing. ICLR 2026. arXiv:2509.21993. doi:10.48550/arXiv.2509.21993.

Kitouni, O., Nolte, N., Bouchacourt, D., Williams, A., Rabbat, M., & Ibrahim, M. (2024). The Factorization Curse: Which Tokens You Predict Underlie the Reversal Curse and More. arXiv:2406.05183. doi:10.48550/arXiv.2406.05183.

Lacroix, T., Usunier, N., & Obozinski, G. (2018). Canonical Tensor Decomposition for Knowledge Base Completion. ICML 2018. arXiv:1806.07297. doi:10.48550/arXiv.1806.07297.

Lampinen, A. K., Chaudhry, A., Chan, S. C. Y., Wild, C., Wan, D., Ku, A., et al. (2025a). On the generalization of language models from

in-context learning and finetuning: a controlled study. arXiv:2505.00661. doi:10.48550/arXiv.2505.00661.

Lampinen, A. K., Engelcke, M., Li, Y., Chaudhry, A., & McClelland, J. L. (2025b). Latent learning: episodic memory complements parametric learning by enabling flexible reuse of experiences. arXiv:2509.16189. doi:10.48550/arXiv.2509.16189.

Li, Z., Zhang, N., Yao, Y., Wang, M., Chen, X., & Chen, H. (2023). Unveiling the Pitfalls of Knowledge Editing for Large Language Models. arXiv:2310.02129. doi:10.48550/arXiv.2310.02129.

Lin, Z., Fu, Z., Liu, K., Xie, L., Lin, B., Wang, W., et al. (2024). Delving into the Reversal Curse: How Far Can Large Language Models Generalize? NeurIPS 2024. arXiv:2410.18808. doi:10.48550/arXiv.2410.18808.

Liu, Z., Michaud, E. J., & Tegmark, M. (2022). Omnigrok: Grokking Beyond Algorithmic Data. arXiv:2210.01117. doi:10.48550/arXiv.2210.01117.

Lv, A., Zhang, K., Xie, S., Tu, Q., Chen, Y., Wen, J.-R., & Yan, R. (2023). An Analysis and Mitigation of the Reversal Curse. arXiv:2311.07468. doi:10.48550/arXiv.2311.07468.

Ma, X., Huang, Y., Zhu, H., & Sojoudi, S. (2026). Breaking the Reversal Curse in Autoregressive Language Models via Identity Bridge. arXiv:2602.02470. doi:10.48550/arXiv.2602.02470.

Merullo, J., Eickhoff, C., & Pavlick, E. (2024). Language Models Implement Simple Word2Vec-style Vector Arithmetic. NAACL 2024. arXiv:2305.16130. doi:10.48550/arXiv.2305.16130.

Nanda, N., Chan, L., Lieberum, T., Smith, J., & Steinhardt, J. (2023). Progress measures for grokking via mechanistic interpretability. ICLR 2023. arXiv:2301.05217. doi:10.48550/arXiv.2301.05217.

Nickel, M., Murphy, K., Tresp, V., & Gabrilovich, E. (2015). A Review of Relational Machine Learning for Knowledge Graphs. Proceedings of the IEEE. arXiv:1503.00759. doi:10.48550/arXiv.1503.00759.

Nie, S., Zhu, F., You, Z., Zhang, X., Ou, J., Hu, J., et al. (2025). Large Language Diffusion Models. arXiv:2502.09992. doi:10.48550/arXiv.2502.09992.

Nishi, K., Ramesh, R., Okawa, M., Khona, M., Tanaka, H., & Lubana, E. S. (2025). Representation Shattering in Transformers: A Synthetic Study with Knowledge Editing. ICML 2025. arXiv:2410.17194. doi:10.48550/arXiv.2410.17194.

Onoe, Y., Zhang, M. J. Q., Padmanabhan, S., Durrett, G., & Choi, E. (2023). Can LMs Learn New Entities from Descriptions? Challenges in Propagating Injected Knowledge. ACL 2023. arXiv:2305.01651. doi:10.48550/arXiv.2305.01651.

Padmanabhan, S., Onoe, Y., Zhang, M. J. Q., Durrett, G., & Choi, E. (2023). Propagating Knowledge Updates to LMs Through Distillation. NeurIPS 2023. arXiv:2306.09306. doi:10.48550/arXiv.2306.09306.

Power, A., Burda, Y., Edwards, H., Babuschkin, I., & Misra, V. (2022). Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets. arXiv:2201.02177. doi:10.48550/arXiv.2201.02177.

Qi, C., Li, B., Hui, B., Wang, B., Li, J., Wu, J., et al. (2023). An Investigation of LLMs' Inefficacy in Understanding Converse Relations. arXiv:2310.05163. doi:10.48550/arXiv.2310.05163.

Qin, J., Zhang, Z., Li, M., Yu, P., & Ji, H. (2024). Why Does New Knowledge Create Messy Ripple Effects in LLMs? arXiv:2407.12828. doi:10.48550/arXiv.2407.12828.

Ravichander, A., Belinkov, Y., & Hovy, E. (2021). Probing the Probing Paradigm: Does Probing Accuracy Entail Task Relevance? EACL 2021. arXiv:2005.00719. doi:10.48550/arXiv.2005.00719.

Sun, Z., Deng, Z.-H., Nie, J.-Y., & Tang, J. (2019). RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. ICLR 2019. arXiv:1902.10197. doi:10.48550/arXiv.1902.10197.

Toutanova, K., & Chen, D. (2015). Observed versus latent features for knowledge base and text inference. Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality. doi:10.18653/v1/W15-4007.

Treutlein, J., Choi, D., Betley, J., Marks, S., Anil, C., Grosse, R., & Evans, O. (2024). Connecting the Dots: LLMs can Infer and Verbalize Latent Structure from Disparate Training Data. arXiv:2406.14546. doi:10.48550/arXiv.2406.14546.

Trouillon, T., Welbl, J., Riedel, S., Gaussier, E., & Bouchard, G. (2016). Complex Embeddings for Simple Link Prediction. ICML 2016. arXiv:1606.06357. doi:10.48550/arXiv.1606.06357.

Varma, V., Shah, R., Kenton, Z., Kramar, J., & Kumar, R. (2023). Explaining grokking through circuit efficiency. arXiv:2309.02390. doi:10.48550/arXiv.2309.02390.

Wang, B., & Sun, H. (2025). Is the Reversal Curse a Binding Problem? Uncovering Limitations of Transformers from a Basic Generalization Failure. arXiv:2504.01928. doi:10.48550/arXiv.2504.01928.

Wang, B., Yue, X., Su, Y., & Sun, H. (2024). Grokked Transformers are Implicit Reasoners: A Mechanistic Journey to the Edge of Generalization. arXiv:2405.15071. doi:10.48550/arXiv.2405.15071.

Wei, Y., Yu, X., Ma, H., Lei, F., Weng, Y., Song, R., et al. (2024). Assessing Knowledge Editing in Language Models via Relation Perspective. arXiv:2311.09053. doi:10.48550/arXiv.2311.09053.

Wu, D., Yang, J., & Wang, K. (2023). Exploring the Reversal Curse and Other Deductive Logical Reasoning in BERT and GPT-Based Large Language Models. arXiv:2312.03633. doi:10.48550/arXiv.2312.03633.

Xu, H.-X., Ma, J.-Y., Ling, Z.-H., Liu, Q., Liu, C., & Gu, J.-C. (2023). Evaluating the Reversal Curse in Model Editing. arXiv:2310.10322. doi:10.48550/arXiv.2310.10322.

Yang, B., Yih, W., He, X., Gao, J., & Deng, L. (2014). Embedding Entities and Relations for Learning and Inference in Knowledge Bases. ICLR 2015. arXiv:1412.6575. doi:10.48550/arXiv.1412.6575.

Zhang, M., Ye, X., Liu, Q., Ren, P., Wu, S., & Chen, Z. (2024). Uncovering Overfitting in Large Language Model Editing. arXiv:2410.07819. doi:10.48550/arXiv.2410.07819.

Zhong, Z., Wu, Z., Manning, C. D., Potts, C., & Chen, D. (2023). MQuAKE: Assessing Knowledge Editing in Language Models via Multi-Hop Questions. EMNLP 2023. arXiv:2305.14795. doi:10.48550/arXiv.2305.14795.

Zhu, H., Huang, B., Zhang, S., Jordan, M., Jiao, J., Tian, Y., & Russell, S. (2024). Towards a Theoretical Understanding of the 'Reversal Curse' via Training Dynamics. arXiv:2405.04669. doi:10.48550/arXiv.2405.04669.

© 2026 Pranay Mahendrakar. Licensed under CC BY 4.0.

DOI: 10.5281/zenodo.24372087