WHEN DOES A PROBE IDENTIFY A REPRESENTATION? SEMANTIC COUNTERFACTUAL VALIDATION IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Decoding accuracy does not show what a probe relies on (Hewitt & Liang, 2019; Ravichander et al., 2021; Belinkov, 2022); for a given probe, no existing test shows whether it tracks the property or its presentation. We propose semantic counterfactual validation, which adds two controls to the evaluation of a probe. The first modifies the presentation of symbolic theories under an exact entailment check, which guarantees that every entailed fact is preserved; a probe of the same type, refitted on the changed presentation with the same training budget, shows whether a failure to transfer leaves the property decodable. The second crosses truth with presentation for a fixed question, using minimally edited twin theories that keep every entity and target concept at the same record positions and counts. All experiments were preregistered with pass or fail conditions except those labelled post-hoc. In type-inheritance theories, a Gemma-3-27B-it probe with AUROC 1.000 falls to 0.892 on a certified rewriting while a refit reaches 0.999, and under reordering a Qwen3-32B probe falls from 0.91 to 0.61 with refits recovering only the original order. On ProofWriter (Tafjord et al., 2021), whose questions need multi-hop reasoning, reordering causes no detectable loss. Twin theories that keep the queried fact true show, in four models, that probes and answers respond to how the records are linked in the original order; a probe trained to separate truth from these edits still fails the certified test. Interchange interventions show that the answer's representation at the last token carries the edit similarly (the two recovery curves differ by at most 0.06) whether or not it changes the truth, and that the statement-site probe direction carries none of this effect. In new theories rendered in random orders, the model answers correctly in every order only when the false alternative is absent from the theory, and there the probes keep their ranking across orders. Decoding accuracy therefore does not establish what a probe tracks, and succeeding the proposed certified test does not establish that the model uses the probe.

1 INTRODUCTION

A probe is a classifier trained on a model's internal activations to predict a property of the input (Belinkov, 2022). Linear probes predict the truth of statements from the activations of large language models with high accuracy (Marks & Tegmark, 2024; Burns et al., 2023), which shows that truth can be read from the activations. This accuracy does not show, however, which feature of the input the probe uses: a probe can reach high accuracy with a feature that correlates with truth only in the way the inputs are written or presented. Since probes are used as evidence of what a model represents, we ask, for a given probe, whether it tracks the property itself or the way the property is presented.

Several results establish that decodability does not identify a representation. Control tasks compare a probe's accuracy with its accuracy on a task that the representation should not support (Hewitt & Liang, 2019). High accuracy can reflect information the model does not use (Ravichander et al., 2021; Elazar et al., 2021), and interchange interventions test whether a representation plays a causal role (Geiger et al., 2021; 2024). Models also learn shortcuts, correlates of the target that hold in their training data (Geirhos et al., 2020), and whether truth probes detect truth is debated (Levinstein & Herrmann, 2025).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

None of these tools decides, for a given probe, whether it tracks a property or the presentation of that property. With natural text, two ingredients are missing. The first is a change of presentation whose effect on meaning is known exactly; without it, a probe that fails to transfer cannot be blamed for the failure. The second is a change of truth that keeps the presentation fixed; without it, a probe that separates true from false inputs may track any difference in wording or position.

We provide both ingredients and use them to test truth probes. Our inputs are small symbolic theories about invented entities and concepts, generated with LatticeForge, a tool built on CG2A, a published generator of conceptual graphs (Faci et al., 2021). A program computes exactly which facts each theory entails. Our method, semantic counterfactual validation, adds two controls to the evaluation of a probe, both defined in Section 2. The first scores the frozen probe after a change of presentation that the program certifies to preserve every entailed fact, and refits a probe of the same family on the changed presentation with the same budget, which tells whether the property is still decodable. The second control crosses truth with presentation for a fixed question. Each true theory has a twin in which two records are edited so that the queried fact becomes false, while every entity and target concept keeps its record positions and counts.

We apply the method to four models; all experiments were preregistered except those labelled post-hoc, and they support four findings.

First, the test separates two kinds of failure. A Gemma-3-27B-it probe with AUROC 1.000 on its training presentation falls to 0.892 on a certified rewriting, while a refit reaches 0.999: truth stays decodable, and only this probe fails to transfer (Section 3.1). Under reordering, the Qwen3-32B probe falls from 0.91 to 0.61, and neither refits nor training in random order recover truth at our budget (Section 3.1). On ProofWriter, a natural-language benchmark whose questions need multi-hop reasoning (Tafjord et al., 2021), the same test passes (Section 3.1).

Second, what a probe tracks depends on its training contrast (Section 3.2). Probes trained to separate true theories from false twins also separate the true theory from twins that keep the queried fact true, as strongly and only in the original record order, in four models. So do the models' answers, and a probe trained to separate truth from these edits still fails the certified test.

Third, interchange interventions show that the representation the answer uses at the last token carries the edit of the links, and that the statement-site probe direction carries none of this effect (Section 3.3).

Fourth, passing the test does not show use (Section 3.4). In new theories rendered in random orders, probes port wherever the model answers correctly in every order, in all four models; these are the questions whose false alternative is absent from the theory. Steering along the probes that pass is specific in none of the four valid runs in these cells, and in one configuration only overall, Qwen3-32B on ProofWriter (Section 3.3).

Our contributions are the following.

- **A method.** Two controls for probes: certified presentation changes with an equal-budget refit, and a crossing of truth with presentation whose twins keep the presentation cues fixed.
- **A diagnosis of truth probes.** In type-inheritance theories, truth probes fail certified changes and track the links between records in the original order, and the models' answers share this sensitivity. What a probe tracks follows its training contrast, and no training contrast we tested gives a probe that passes the certified test on these theories.
- **What a pass shows.** The test passes on multi-hop questions in ProofWriter and on presence questions in our theories. A pass shows a probe that holds across these presentations, and it does not show that the answer uses the probe.

A failed test does not show that a model lacks a representation of truth; the refit tells whether truth remains decodable.

2 Setting and definitions

This section defines every object used in Section 3.

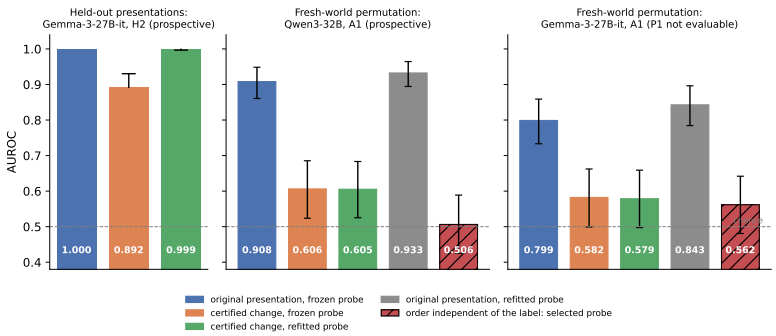

Figure 1: The two failure outcomes of the portability test, both from preregistered experiments. Left, the test of held-out presentations (Section 3.1): the frozen Gemma-3-27B-it probe loses part of its ranking on the rewriting H2, in which the queried fact needs three inference steps (the loss is at least 0.070 at the preregistered one-sided level, adjusted for multiple comparisons), while a refitted probe of the same family reaches AUROC 0.999 (lower bound 0.997): the property remains decodable along another direction. Middle and right, the fresh-theory reordering test (Section 3.1; 192 TEST theories): after a certified reordering of the records, the frozen probes fall from 0.91 to 0.61 in Qwen3-32B (paired upper bound of the change -0.21) and from 0.80 to 0.58 in Gemma-3-27B-it, whose probe misses the preregistered accuracy check on fresh theories in the original order; refitted probes, trained on 88 to 104 theories, reach 0.93 and 0.84 in the original order but only 0.61 and 0.58 after reordering. Whiskers: 95% bootstrap intervals over theories (middle, right) and one-sided bounds (left). Hatched: probes trained on theories whose record order was drawn independently of the label, scored on reordered TEST theories. Original-presentation AUROCs are not comparable across the two experiments, which use different theories.

A theory is a set of records of two kinds. A type record $[\mathrm{type}, e, c]$ states that entity e is of type c, and a parent record $[\mathrm{parent}, c_1, c_2]$ states that every entity of type c_1 is also of type c_2. A theory entails every type fact reachable from an entity's types through parent records, its closure, which a deterministic program, the oracle, computes exactly. Names are fresh random strings, so that no name carries a meaning and none recurs across data splits. Most experiments use eight-record theories in which E and X are typed by A_0 and B_0, and two chains of parent records lead from A_0 through A_1 and A_2 to T and from B_0 through B_1 and B_2 to U. The true theory W^+ entails type(E,T) after three inference steps, and we call E, X, T and U the role symbols. A false question either names a concept absent from the theory or is a structural negative, whose queried concept is reachable from X but not from E. Structural negatives come in three tiers, which exchange the concepts of the two type records (NEG_PREFIX0), the destinations of the second parent record of each chain (NEG_PREFIX1), or those of the last parent record (NEG_PREFIX2).

A presentation is the text that renders a theory. A presentation change permutes the records, whose canonical order is the generator's, or rewrites them into another set with the same closure; on ProofWriter it permutes the sentences of a theory. The oracle checks every change before any model runs, and the closure must be unchanged. We call this check a certificate. A twin of W^+ is a theory obtained by editing records while every role symbol keeps its record positions and counts. The false twin W^- makes type(E,T) false; the truth-preserving twins of Section 3.2 keep it true. Twins are minimal pairs, as interchange interventions require (Geiger et al., 2021); the constraint on the role symbols is our addition. The crossing's twin exchanges the destinations of records 3 and 6 (numbered from 0), which makes it the NEG_PREFIX1 negative of W^+; its certificate checks both properties.

We study Qwen3-32B and Gemma-3-27B-it and, in five experiments, Qwen3-8B and Gemma-3-12B-it, at fixed revisions, in bfloat16 and without generated reasoning (Appendix G). A statement prompt ends with the queried fact as a sentence and asks the model not to answer. A forced-choice prompt ends with the question whether the fact is entailed and asks for the letter A or B. Under the response mapping M0, A means entailed; M1 swaps the letters. The answer gap is the logit of the letter that means entailed minus the logit of the other. Activations are the outputs of decoder blocks at one token position, the site: the final period of a statement, the question mark of a forced-choice

prompt (the question-mark site), or the last token. A window averages three consecutive blocks, written W00 to W07 for each model. A probe family turns activations into a score, and a probe is a trained member of a family with its window and regularization fixed. We use four established linear families: the mass-mean direction (Marks & Tegmark, 2024), logistic regression, a linear SVM, and CCS (Burns et al., 2023); the controls apply to any probe. Scores are summarized by the AUROC and, for matched pairs, by the paired ranking, the fraction of pairs in which the true member scores higher, ties counting one half. IID evaluation scores held-out cases drawn from the training presentation. Intervals are 95% bootstrap intervals over theories or blocks of related cases, and one-sided bounds serve directional criteria.

We distinguish three levels of evidence, a distinction of ours that names the gap control tasks point to (Hewitt & Liang, 2019; Ravichander et al., 2021): a property can be decodable within a presentation (L1), a frozen probe can keep its ranking across certified presentations (L2), and its scores can remain comparable (L3). The experiments test L2, which we call portability; a portability failure is a loss that meets the preregistered criterion of the experiment. Our first control, the equal-budget refit, trains a probe of the same family, with the same selection procedure and number of cases, on the changed presentation; recoverability is demonstrated when this refit reaches the preregistered level, and is not established otherwise. Our second control, the crossing, renders W^+ and its twin in canonical order and in one certified reordering shared by both, and asks the same question of the four renderings, a quartet. It reports the paired ranking of W^+ over W^- in each order and the interaction of truth with order, the change of the mean score difference between W^+ and W^- from canonical to reordered. The positional cue is a baseline computed from the records alone: a logistic regression on which record positions link to which, and on where E and T occur. Since the theory determines its label in every order, a records-only model allowed to compute reachability would be perfect; the cue is therefore restricted to features indexed by position.

Steering adds a multiple $\alpha\sigma u$ of a probe's unit direction u to the outputs of the blocks of its window at every position, where σ is the standard deviation of the projections on TRAIN cases; interventions along truth directions follow Marks & Tegmark (2024) and Li et al. (2023). Interchange interventions replace the output of one block at one site in one input by its value in another input (Geiger et al., 2021; 2024; 2025). An experiment is preregistered when its hypotheses, probes, estimands and decision rules were hashed and logged before any model output of that experiment was computed. Registered readings are printed in capitals, and their rules are given with each protocol in the appendices. A positive control checks that the design separates true from false where it must, and its failure makes the main outcome not evaluable. A post-hoc analysis is labelled as such, and its rationale was logged before computation. No experiment was rerun to improve its outcome, and TRAIN, VALIDATION and TEST splits are disjoint in theories and names.

3 Experiments

The experiments follow the two controls. Section 3.1 changes the presentation of fixed worlds and asks whether frozen probes keep their ranking. Section 3.2 changes the truth while keeping the presentation fixed, to find what the probes track. Section 3.3 asks whether the answer uses what the probes read. Section 3.4 applies the test where the model's competence does not depend on the order.

3.1 Changing the presentation of a fixed world

This subsection asks whether a probe that decodes truth keeps its ranking when the presentation changes and the meaning provably does not, and, if it fails, whether truth is still decodable.

The first experiment scores held-out rewritings of the same worlds. Probes are trained on the difference between the activations of a positive and a negative statement of the queried fact, selected on seen presentation families, then frozen and scored on held-out families. The seen families state the queried fact directly (S0 and S1, two serializations) or place it one inference step away (S2). The held-out families place it two steps away (H1), keep only the records of a three-step derivation (H2), state the whole closure in shuffled order (H3), or state the fact directly in a third serialization (H4). Each of nine systems per model is tested only if it passes a preregistered IID gate that includes a selectivity control, refits with the TRAIN labels of whole blocks flipped (a block-level analogue of

Model	Frozen probe	Canonical [95% CI]	Permuted [95% CI]	Change [95% CI]	Portability	Refit, permuted [95% CI]	Refit, canonical [95% CI]
Qwen3-32B	IID, W05 SVM C=0.01	0.908 [0.861, 0.949]	0.606 [0.524, 0.685]	-0.302 [-0.388, -0.212]	success	0.605 [0.525, 0.683]	0.933 [0.894, 0.964]
Qwen3-32B	DIVERSE, W05 SVM C=1	0.909 [0.862, 0.949]	0.606 [0.525, 0.685]	-0.303 [-0.390, -0.215]	success	0.604 [0.525, 0.684]	0.933 [0.895, 0.965]
Gemma-3-27B-it	IID, W05 LogReg	0.799 [0.733, 0.859]	0.582 [0.499, 0.662]	-0.216 [-0.315, -0.118]	not evaluable	0.579 [0.497, 0.659]	0.843 [0.784, 0.896]
Gemma-3-27B-it	DIVERSE, W06 SVM	0.806 [0.743, 0.863]	0.573 [0.493, 0.654]	-0.233 [-0.331, -0.133]	not evaluable	0.566 [0.486, 0.646]	0.844 [0.787, 0.895]
Qwen3-32B	answer gap (post-hoc)	0.912 [0.872, 0.947]	0.590 [0.509, 0.668]	-0.323 [-0.408, -0.239]	—	—	—
Gemma-3-27B-it	answer gap (post-hoc)	0.719 [0.646, 0.788]	0.509 [0.426, 0.589]	-0.209 [-0.320, -0.101]	—	—	—

Table 1: The fresh-world permutation replication, reordering panel (preregistered; 192 TEST cases). Frozen probes from the selection-rule development, scored on fresh canonical renderings and on certified permutations of the same cases; the last two rows give the model's own answer gap on the same cases (post-hoc reanalysis of the recorded logits). IID-selected and DIVERSE name the two selection rules preregistered as primary (DIVERSE, the matched control of the counterfactual-audit rule, selects on an audit set of the same size whose cases vary in surface form, context density and graph topology instead of being certified transformations); the two Qwen rows are one probe (Spearman $\rho = 0.999$ between their TEST scores, post-hoc). Intervals from 10,000 case bootstraps, paired for the change. A loss is detected when the upper bound of the change lies below -0.10 and the canonical one-sided lower bound is at least 0.80; the Gemma probes miss the latter, hence not evaluable. Refits use the same family, window and regularization strength, cross-fitted over two folds assigned by hashing world names (training sets of 88 and 104 cases); recoverability is established when the one-sided lower bound is at least 0.75 (permuted: 0.538, 0.538, 0.512, 0.498; canonical: all above). Learning curves at 24, 48 and 96 training cases: canonical 0.84, 0.87, 0.92 (Qwen) and 0.72, 0.67, 0.85 (Gemma); permuted 0.55, 0.52, 0.62 and 0.51, 0.54, 0.59.

control tasks; Hewitt & Liang, 2019), and it fails L2 when the adjusted upper bound of its AUROC change lies below -0.05 (Appendix O).

The Gemma-3-27B-it logistic probe of the mixed-unpaired regime passes the gate (IID AUROC 1.000, selectivity $p = 0.003$). It keeps 1.000 on H1 and H4 and 0.945 on H3, but falls to 0.892 on H2, while a refit on H2 reaches 0.999 (Figure 1). The registered primary required this pattern in both models and was not met, because no Qwen3-32B system passed the gate: 16 of 18 systems reach IID AUROC ≥ 0.999 with selectivity p from 0.06 to 0.50, since the block-flip null only randomizes the sign of the truth direction (post-hoc diagnostic, Table 18). The H2 loss comes from one negative tier: the separation of NEG_PREFIX0 falls to 0.36 [0.19, 0.52], against 1.86 and 2.00 for the other two tiers (Table 20). This tier keeps the parent records of the true world and changes only which chain the entity enters, so the loss concerns the negatives whose parent records match the true world. This is the first outcome that the test separates: the frozen probe fails while the property stays linearly decodable, and IID evaluation did not reveal it.

The reordering effect was first seen post-hoc, in a development experiment on probe-selection rules (Appendix L), where frozen probes fell from about 0.90 to 0.54–0.58 AUROC under a certified permutation of the records. The fresh-world replication then tested it prospectively, on 576 new worlds (TRAIN 288, VALIDATION 96, TEST 192). Its primary outcome is portability under one certified permutation (Table 1). The frozen Qwen3-32B probes fall from 0.91 to 0.61, with a paired upper bound of -0.21, below the criterion of -0.10. The Gemma probes reach only 0.80 in canonical order, below the positive control, so Gemma is not evaluable for this outcome ($0.80 \rightarrow 0.58$ descriptively). Refits cross-fitted on training sets of 88 and 104 cases reach 0.93 (Qwen) and 0.84 (Gemma) in canonical order, and 0.61 and 0.58 after the permutation, where recoverability is not established. Probes trained on worlds whose record order is independent of the label reach only 0.51 and 0.56 (Table 14): at this budget, truth cannot be decoded from these activations once the canonical layout is removed. The model's own answer shares the collapse. Its registered behavioral secondary reads DEGRADED in both models. In a post-hoc reanalysis of the recorded logits, whose plan was logged before computation, the AUROC of the answer gap falls from 0.912 to 0.590 in Qwen3-32B and from 0.719 to 0.509 in Gemma-3-27B-it. This is the second outcome: a collapse that no refit recovers at this budget and that the model's own answer shares.

ProofWriter tests whether this collapse is general. Permuting the sentences of each theory leaves every entailment unchanged. The frozen direction at the question-mark site shows no detectable loss (Appendix I; the intervals include 0): its row AUROC moves from 0.975 to 0.971 in Qwen and from 0.891 to 0.902 in Gemma, a direction relearned on permuted theories is almost identical (cosine 0.98 and 0.99), and accuracy does not drop ($0.72 \rightarrow 0.73$ and $0.70 \rightarrow 0.73$). The collapse is therefore specific to the type-inheritance theories. Asking true and false questions about the same

type-inheritance theory, as ProofWriter does, does not remove it: in a preregistered control, a post-hoc predictor using record positions alone reaches 0.89 in canonical order, and a same-family SVM falls from 0.93 to 0.52 under permutation in Qwen (Appendix J). Questions about different theories are therefore not the reason why the collapse is specific to the type-inheritance theories, and the crossing removes the remaining link between positions and labels.

3.2 Changing the truth while holding the presentation fixed

The experiments above show that the probes fail to transfer. The crossing and its twins show what the probes track, by changing the truth of the queried fact while keeping the presentation fixed.

The crossing (preregistered) uses 576 new quartets. In each, the false twin W^- exchanges the destinations of records 3 and 6 of the true world W^+; every symbol keeps its frequency, and E, X, T and U keep their record positions. Order B applies one certified permutation per quartet to both theories. On 384 TEST quartets, the registered Qwen3-32B probe, the linear SVM of the pairing control, passes the positive control and ranks W^+ above W^- in 0.987 [0.974, 0.997] of quartets in canonical order and in 0.495 [0.445, 0.544] after permutation (interaction -1.46 [-1.54, -1.37]; reading PRESENTATION-CONDITIONED). The Gemma probe misses the positive control (Appendix B). Probes trained on the positive statement of the crossing's canonical TRAIN quartets, the in-panel probes (post-hoc, frozen before any TEST or permuted activation was read), do the same in all four models (Table 4, Appendix A). So does the positional cue (1.000 and 0.483), whereas no cue that ignores record positions can separate the twins without computing reachability. The probes' sensitivity to the edit thus disappears when the same theory is reordered, but the crossing cannot tell whether they track the truth or the edit, since the edit changes both.

The truth-preserving twins (preregistered) separate the edit from the truth. Each of 576 new worlds has, besides W^+ and W^-, three twins in which type(E,T) stays entailed. The rewire twin W^r redirects the same records 3 and 6, so that T is reached one record later. The frequency-matched twin W^f also edits record 5, so that every symbol occurs as often as in W^+. The skip twin W^s leaves records 3 and 6 unchanged and redirects records 2 and 5, which shortens the proof by one record. A probe of the truth would separate only W^-, and a probe of the wiring of records 3 and 6 would separate W^-, W^r and W^f but not W^s (Table 5); our stated prediction was the wiring account for the in-panel probes.

The in-panel probes separate every twin in canonical order in all four models (Table 2). Their contrast, the mean score difference between W^+ and a twin in units of the probe's TRAIN standard deviation, for W^r and W^f is 0.94 to 1.07 times their contrast for W^-, so the truth flip adds nothing to a truth-preserving rewiring; the skip twin receives 0.17 to 0.39 of that contrast. After reordering, every ranking falls to between 0.35 and 0.58 and every contrast to about zero. The registered reading is UNCLASSIFIED in every model, and our prediction is not confirmed. The registered secondary refits the probe with all five cells in training and the truth-preserving twins labelled true. This probe learns to separate truth from the edits in canonical order: its contrast for W^r and W^f falls to 0.06 to 0.25 of its contrast for W^-. It still fails the certified test, with a W^- ranking of 0.54 or less after reordering and contrasts near zero. What a probe tracks therefore follows its training contrast, and even a probe trained to ignore the wiring does not track a truth that holds across presentations.

The model's own answer, run on the same twins with forced-choice prompts (preregistered), responds to the edits as the in-panel probes do. In canonical order its contrast for W^r and W^f is 0.76 to 1.30 times its contrast for W^-, and its rankings are 0.91 to 1.00; after reordering they fall to between 0.46 and 0.63. The registered reading is UNCLASSIFIED in three models and WIRING ORDER ROBUST in Gemma-3-27B-it, whose permuted W^- ranking has an upper bound of 0.6003, just above the rule's cut at 0.60, and whose answer ties the skip twin (0.475).

3.3 From decodability to use

Decoding cannot show whether the model uses what a probe reads. Two interventions test it: steering along the probes, and interchange interventions that locate the representation the answer uses.

Steering (Section 2) is applied during the forced-choice prompts, with a random orthogonal direction as control. The registered estimand, the change of the answer gap signed by the correct answer and averaged over true and false rows, cancels by construction for a truth direction, since a push toward

		ranking $P[s(W^+) > s(\cdot)]$				contrast (SD units)			
Model	Scorer	W^-	W^r	W^f	W^s	W^-	W^r	W^f	W^s
Qwen3-32B	in-panel probe	1.000	1.000	1.000	0.990	1.95	1.96	1.91	0.36
	all-cells refit	1.000	0.802	0.922	0.143	1.89	0.30	0.45	−0.30
	answer	0.951	0.986	0.906	0.863	1.32	1.61	1.00	0.62
Gemma-3-27B-it	in-panel probe	1.000	1.000	1.000	1.000	1.96	1.99	1.94	0.76
	all-cells refit	1.000	0.620	0.872	0.445	1.92	0.12	0.36	−0.06
	answer	1.000	0.995	1.000	0.475	1.23	1.11	1.23	−0.03
Qwen3-8B	in-panel probe	1.000	1.000	1.000	0.995	1.90	2.03	1.79	0.66
	all-cells refit	1.000	0.857	0.943	0.721	1.91	0.35	0.47	0.16
	answer	0.987	0.990	0.941	0.685	1.35	1.76	1.03	0.24
Gemma-3-12B-it	in-panel probe	1.000	1.000	1.000	0.992	1.90	2.00	1.85	0.33
	all-cells refit	1.000	0.648	0.844	0.570	1.88	0.16	0.39	0.04
	answer	0.999	0.997	1.000	0.973	1.53	1.58	1.55	0.93
records only	positional cue	1.000	1.000	1.000	0.500				

Table 2: The truth-preserving twins, canonical order, preregistered (384 TEST worlds). W^- makes the queried fact false; W^r, W^f and W^s keep it true. A probe of the truth would separate only W^-; a probe of the wiring of records 3 and 6 would leave W^s at chance, as the positional cue does (all signatures in Table 5). The in-panel probes are trained on W^+ against W^-; the all-cells refit is trained on the five cells with the truth-preserving twins labelled true; the answer is the mapping-corrected answer gap. Contrast: mean score difference between W^+ and the twin, in units of the scorer's TRAIN standard deviation. After reordering, every ranking lies between 0.22 and 0.63 and every contrast between -0.26 and 0.20 (the probes after reordering in Table 6, and the answer's values, readings and gates in Table 7, Appendix A).

"entailed" raises it on true rows and lowers it on false rows equally. We therefore report a directional estimand fixed before computation, the change of the answer gap between $\alpha = -3$ and $+3$ (post-hoc; Table 8, Appendix A). On ProofWriter, where the probe passes the test, steering moves the Qwen3-32B answer by $+8.4$ logits in canonical order and $+8.3$ after permutation, monotonically and about 9 logits more than the random control. In Gemma-3-27B-it the random direction moves the answer as well, so the effect is not specific (Appendix N). In the type-inheritance theories, where the probes fail the test, steering changes the answer gap by -0.07 to -0.25 logits across the frozen probes (-0.07 to -0.13 for those of Table 8), some fifty times less than on ProofWriter, although the answer separates true from false theories in canonical order.

Interchange interventions locate the representation that the answer uses (preregistered). The twins W^- and W^f are token-matched to W^+, so the block output of a twin at one site can be replaced by that of W^+, and the recovery ratio, a standard patching metric (Zhang & Nanda, 2024), measures the share of the answer difference that moves with it. The validity gates, which check that unpatched runs reproduce the recorded logits and that patching a run with its own activations changes nothing, pass in all four models; the reproduction gate was amended after it failed in two models on raw logits, before any estimate was computed (Appendix B). A first registered panel selected the final token at the last block, whose output fixes the logits, so its readings are uninformative (a design error of its selection rule); a second panel, logged before computation, takes, for each question-side site (the queried concept, the question mark and the last token), the earliest block below the last whose recovery on TRAIN worlds has a one-sided lower bound of at least 0.5.

The last token is the only question-side site that reaches this criterion, at block 52 of 64 in Qwen3-32B, 36 of 62 in Gemma-3-27B-it, 24 of 36 in Qwen3-8B and 28 of 48 in Gemma-3-12B-it. At every block, the recovery for the false twin and the recovery for the truth-preserving twin differ by at most 0.06 (Figure 2). At the chosen block they are 0.900 and 0.907 in Qwen3-32B, 0.546 and 0.546 in Gemma-3-27B-it, 0.946 and 0.933 in Qwen3-8B, and 0.970 and 0.980 in Gemma-3-12B-it. The representation the answer uses at the last token therefore carries the edit of the links identically whether or not the edit changes the truth (reading REWIRING-CODED, as predicted). After reordering, the answer difference shrinks to between -0.07 and $+0.41$ logits, and the replacement moves the answer by at most 0.21 logits (PRESENTATION-CONDITIONED). The unit direction

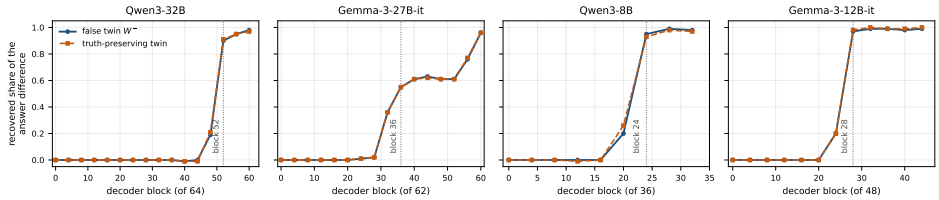

Figure 2: Where the answer's dependence on the edit is carried, preregistered. Share of the answer difference between the true theory and a twin that is recovered when the twin's last-token activation at one decoder block is replaced by that of the true theory (384 TEST theories, both response mappings, canonical order). Solid: the false twin W^-; dashed: the frequency-matched twin, which keeps the queried fact true. Dotted line: the block chosen on TRAIN theories, the earliest block, excluding the last, whose recovery has a one-sided lower bound of at least 0.5. The two curves coincide at every block, so the representation the answer uses carries the edit of the links whether or not it changes the truth. Values as logged in the result files.

of the in-panel probe, estimated at the statement site of statement prompts, carries none of this effect at the answer site (OFF PATHWAY); the interchange does not test what the answer does with the statement-site representation.

3.4 A CONFIGURATION THE MODEL SOLVES IN EVERY ORDER

The failures above come from theories whose canonical layout reveals the label. A preregistered panel asks whether the test passes where competence does not depend on the order, and whether a passing probe is used, on 384 new worlds. Each world gets seven questions about E: three true ones, whether E is of type A_1, A_2 or T (one, two or three inference steps); three structural false ones, about the concepts at the same depth of the other chain; and one about a concept absent from the theory. Pairing each true question with the false question of the same depth, or with the absent concept, gives six cells, and in each cell both questions concern the same theory. Every theory is rendered in canonical order and in two random orders, P and Q, drawn independently of any question, and in each cell the panel measures the answer's AUROC and trains a probe in P scored in Q, and one in canonical order scored in P. A cell is competent when the answer's lower bound is at least 0.80 in both random orders; the primary reading requires that the competent cells port and that no other cell does, and our prediction was this reading in all four models.

The only competent cells are those whose false alternative is the absent concept: there the answer reaches 0.81 to 0.99 in both random orders, and the probes are perfect in every order and port perfectly (Table 3). In these cells the presence of the queried name decides the answer in any order, so the panel shows that the test passes a presence detector; ProofWriter is where it passes on multi-hop questions, although the model's accuracy on entailed questions at depths 2 and 3 is only 0.62 to 0.72 in either order. With structural negatives no model is competent in random order, and probes trained in canonical order fall from 1.000 to between 0.47 and 0.53, so the reordering collapse replicates in all four models and at all three depths. The primary reading holds in Qwen3-32B, Qwen3-8B and Gemma-3-12B-it, and is VIOLATED BY PORT in Gemma-3-27B-it, where three cells port although the answer misses the competence criterion (at depth 1 with a structural negative, the probe reaches 0.84 and 0.87 while the answer reaches 0.64 and 0.61).

A registered second stage steered along the probes of the ten cells that are competent and port. No cell reads SPECIFIC. The six Qwen runs are invalid by the registered reproduction gate (sign agreement below the required 99% in at least one order), and in the four valid Gemma runs the random direction moves the answer as well (-3.4 to $+6.5$ logits, against $+2.7$ to $+10.2$ for the probe).

4 RELATED WORK

Linear structure related to truth has been found across datasets and models (Burns et al., 2023; Marks & Tegmark, 2024; Bürger et al., 2024). Propositional probes read facts stated in context (Feng et al.,

Model	negative	answer AUROC, random orders P/Q depth 1	2	3	probe trained in P, scored in Q 1	2	3	probe trained canonically, scored in P 1	2	3
Qwen3-32B	structural	0.66/0.71	0.52/0.57	0.52/0.53	0.820	0.544	0.518	0.481	0.491	0.528
	fresh	0.95/0.96	0.93/0.93	0.95/0.94	1.000	1.000	1.000	1.000	1.000	1.000
Gemma-3-27B-it	structural	0.64/0.61	0.52/0.52	0.50/0.51	0.870†	0.565	0.510	0.472	0.511	0.524
	fresh	0.92/0.91	0.83/0.83	0.81/0.81	1.000	1.000†	1.000†	1.000	1.000	1.000
Qwen3-8B	structural	0.62/0.60	0.52/0.51	0.48/0.52	0.636	0.512	0.513	0.476	0.508	0.511
	fresh	0.91/0.93	0.86/0.87	0.94/0.97	1.000	1.000	1.000	1.000	1.000	1.000
Gemma-3-12B-it	structural	0.61/0.63	0.49/0.49	0.47/0.50	0.608	0.502	0.512	0.478	0.522	0.519
	fresh	0.99/0.99	0.98/0.96	0.93/0.92	1.000	1.000	1.000	1.000	1.000	1.000

Table 3: Competence and portability across six cells, preregistered (192 TEST worlds per cell; the true and the false question of a pair concern the same theory). A cell is competent when the answer's one-sided lower bound is at least 0.80 in both random orders: every fresh cell except depths 2 and 3 in Gemma-3-27B-it, and no structural cell. Probes trained in canonical order reach 1.000 there in every cell and model. †: the three cells in which a probe ports although the answer is not competent (Gemma-3-27B-it, reading VIOLATED BY PORT); the primary reading COMPETENCE PREDICTS PORTABILITY holds in the three other models. Probes trained in P reach, on TEST cases in P, the values of the middle block within 0.06.

2025), and truth and hallucination probes generalize unevenly across tasks (Orgad et al., 2025). Truth directions depend on context (Schouten et al., 2025; Adarsh et al., 2026) and have been tested for consistency across logical transformations (Bao et al., 2025). Refitting on the changed condition is common in these studies (Schouten et al., 2025); our refit adds a certificate that the changed presentation keeps every entailed fact, and an equal training budget, and the crossing with its twins tells whether the probe responds to the truth or to the presentation.

Interventions on subspaces can act through pathways the model does not use (Makelov et al., 2024), which bounds what steering along a probe can show. Reasoning on fictional ontologies and stated rules has been evaluated behaviorally (Saparov & He, 2023; Clark et al., 2020; Tafjord et al., 2021), and latent multi-hop reasoning and logical circuits have been studied inside models (Yang et al., 2024; Biran et al., 2024; Yang et al., 2025; Hong et al., 2025; Fang et al., 2026). Our theories add exact certificates and opaque names, so that what is entailed is known and held fixed while the presentation changes. Position- and order-based binding mechanisms (Feng & Steinhardt, 2024; Dai et al., 2024) are the closest existing account of why probes and answers depend on record order. The position and order of information in the context change behavior (Liu et al., 2024; Chen et al., 2024), and so do its serialization (Fatemi et al., 2024) and its repetition (Xu et al., 2024); our permutation results are a counterpart at the level of representations.

5 Discussion and limitations

The test separates three experimental outcomes. A frozen probe can fail while the property stays decodable (H2), fail with decodability not established (reordering in the type-inheritance theories), or pass (ProofWriter, and the presence cells). A pass does not certify a semantic representation, since a shortcut can be invariant to a given change. Four further preregistered experiments show that logically equivalent theories differ in how accessible their facts are to the model, which is why the presentation must be controlled (Appendix C).

The type-inheritance results are controlled shortcut learning, as found for models trained on rule-reasoning data (Zhang et al., 2023): the canonical layout correlates with the label, and the test detects the shortcut where IID validation does not.

The rewriting counterexample rests on one Gemma probe and the fresh-world collapse on Qwen3-32B, while the later panels replicate in four models. A probe trained at the answer site remains to be examined under the certified test, and the scope is bounded by our synthetic theories, linear probes and four open models of two families.

In conclusion, IID accuracy does not show what a probe tracks, and certified presentation changes with a refit and a crossing of truth with presentation do; on truth probes, they show that what a probe tracks follows its training contrast.

AI USE STATEMENT

We used generative AI tools to write and debug code. We also used them to redraft from notes and draft, proofread the text for typos, misspellings, wording and grammar, to refine its style, to look for inconsistencies, to harmonize naming conventions, and to help verify the anonymization of the supplementary material. We did not use them to generate data, run experiments or produce results; all results come from the authors' pipelines on the authors' infrastructure. The authors reviewed all AI-assisted code, tested it on synthetic data and checked it against previously recorded results before use, and reviewed and edited all AI-assisted text. We take responsibility for the full content of this work.

ETHICS STATEMENT

This work uses only synthetic symbolic worlds with freshly generated opaque symbols and one public benchmark (ProofWriter); no human subjects, personal data, or sensitive content are involved. The methodological point, that decoding accuracy can overstate what a model represents, bears on the reliability of interpretability claims used in auditing and safety arguments; we see no direct misuse risk beyond that of interpretability research in general.

REPRODUCIBILITY STATEMENT

All primary analyses were preregistered with PASS/FAIL verdicts and frozen before the corresponding forward passes; failed criteria are reported in the claim-to-evidence map (Appendix B). The order-variable diagnostic analysis plan, presentation-variable schema, fold assignment, and per-model data files were hashed before computation (plan SHA256 in the provenance files of the supplementary material); the diagnostic verifies these hashes and reproduces the previously recorded source and permuted AUROCs before any inference. Model revisions, window indices, and forward counts are listed in Appendix G. The supplementary material contains, for every experiment reported, the anonymized code, the frozen protocols and analysis plans, the worlds and certificates from which every prompt can be re-rendered, and the result files, with a crosswalk from each released file to the SHA256 logged during the project; for three forced-choice panels (the fresh-world replication, the truth-preserving twins and the competence panel) it also contains the prompt panels and the answer logits. The complete code, prompt panels and model outputs of the project, including its development runs, will be released publicly upon acceptance.

REFERENCES

Shivam Adarsh, Maria Maistro, and Christina Lioma. How Context Shapes Truth: Geometric Transformations of Statement-level Truth Representations in LLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 36570–36590, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1695. URL https://aclanthology.org/2026.acl-long.1695.

Yuntai Bao, Xuhong Zhang, Tianyu Du, Xinkui Zhao, Zhengwen Feng, Hao Peng, and Jianwei Yin. Probing the Geometry of Truth: Consistency and Generalization of Truth Directions in LLMs Across Logical Transformations and Question Answering Tasks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 682–700, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.38. URL https://aclanthology.org/2025.findings-acl.38.

Yonatan Belinkov. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, 48(1):207–219, April 2022. ISSN 0891-2017, 1530-9312. doi: 10.1162/coli_a_00422. URL https://direct.mit.edu/coli/article/48/1/207/107571/Probing-Classifiers-Promises-Shortcomings-and.

Eden Biran, Daniela Gottesman, Sohee Yang, Mor Geva, and Amir Globerson. Hopping too late: Exploring the limitations of large language models on multi-hop queries. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 14113–14130, 2024.

Lennart Bürger, Fred A. Hamprecht, and Boaz Nadler. Truth is universal: Robust detection of lies in LLMs. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*, 2023.

Xinyun Chen, Ryan A. Chi, Xuezhi Wang, and Denny Zhou. Premise order matters in reasoning with large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 2024.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. Transformers as Soft Reasoners over Language. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pp. 3882–3890, Yokohama, Japan, July 2020. International Joint Conferences on Artificial Intelligence Organization. ISBN 978-0-9992411-6-5. doi: 10.24963/ijcai.2020/537. URL https://www.ijcai.org/proceedings/2020/537.

Qin Dai, Benjamin Heinzerling, and Kentaro Inui. Representational Analysis of Binding in Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17468–17493, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.967. URL https://aclanthology.org/2024.emnlp-main.967.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, March 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00359. URL https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00359/98091/Amnesic-Probing-Behavioral-Explanation-with.

Adam Faci, Marie-Jeanne Lesot, and Claire Laudy. CG2A: Conceptual Graphs Generation Algorithm. In *Joint Proceedings of the 19th World Congress of the International Fuzzy Systems Association (IFSA), the 12th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT), and the 11th International Summer School on Aggregation Operators (AGOP)*, pp. 63–70. Atlantis Press, August 2021. ISBN 978-94-6239-423-0. doi: 10.2991/asum.k.210827.009. URL https://www.atlantis-press.com/proceedings/ifsa-eusflat-agop-21/125960379.

Feihao Fang, My T. Thai, and Yuanyuan Lei. Discovering a Shared Logical Subspace: Steering LLM Logical Reasoning via Alignment of Natural-Language and Symbolic Views. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 38936–38956, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1806. URL https://aclanthology.org/2026.acl-long.1806.

Bahare Fatemi, Jonathan Halcrow, and Bryan Perozzi. Talk like a graph: Encoding graphs for large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Jiahai Feng and Jacob Steinhardt. How do language models bind entities in context? In *The Twelfth International Conference on Learning Representations*, 2024.

Jiahai Feng, Stuart Russell, and Jacob Steinhardt. Monitoring latent world states in language models with propositional probes. In *The Thirteenth International Conference on Learning Representations*, 2025.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9574–9586, 2021.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 160–187, 2024.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, November 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00257-z. URL https://www.nature.com/articles/s42256-020-00257-z.

John Hewitt and Percy Liang. Designing and Interpreting Probes with Control Tasks. 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)*, pp. 2733–2743, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275. URL https://www.aclweb.org/anthology/D19-1275.

Guan Zhe Hong, Nishanth Dikkala, Enming Luo, Cyrus Rashtchian, Xin Wang, and Rina Panigrahy. A implies B: Circuit analysis in LLMs for propositional logical reasoning. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Benjamin A. Levinstein and Daniel A. Herrmann. Still no lie detector for language models: probing empirical and conceptual roadblocks. *Philosophical Studies*, 182(7):1539–1565, July 2025. ISSN 0031-8116, 1573-0883. doi: 10.1007/s11098-023-02094-3. URL https://link.springer.com/10.1007/s11098-023-02094-3.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, February 2024. ISSN 2307-387X. doi: 10.1162/tacl_a_00638. URL https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00638/119630/Lost-in-the-Middle-How-Language-Models-Use-Long.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *The Twelfth International Conference on Learning Representations*, 2024.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*, 2024.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *The Thirteenth International Conference on Learning Representations*, 2025.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the Probing Paradigm: Does Probing Accuracy Entail Task Relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3363–3377, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.295. URL https://aclanthology.org/2021.eacl-main.295.

Abulhair Saparov and He He. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *The Eleventh International Conference on Learning Representations*, 2023.

Stefan F. Schouten, Peter Bloem, Ilia Markov, and Piek Vossen. Truth-value judgment in language models: 'truth directions' are context sensitive. In *Second Conference on Language Modeling*, 2025.

Model	Canonical AUROC	Ranking, canonical [95% CI]	Ranking, permuted [95% CI]	Interaction [95% CI]	Reading
Qwen3-32B	1.000	1.000 [1.000, 1.000]	0.542 [0.492, 0.591]	-1.96 [-2.01, -1.92]	presentation-conditioned
Gemma-3-27B-it	1.000	1.000 [1.000, 1.000]	0.495 [0.443, 0.544]	-2.53 [-2.60, -2.47]	presentation-conditioned
Qwen3-8B	1.000	1.000 [1.000, 1.000]	0.508 [0.458, 0.557]	-1.20 [-1.23, -1.17]	presentation-conditioned
Gemma-3-12B-it	1.000	1.000 [1.000, 1.000]	0.516 [0.466, 0.565]	-2.50 [-2.56, -2.43]	presentation-conditioned
Position-indexed cue, no activations	1.000	1.000 [1.000, 1.000]	0.486 [0.461, 0.510]	—	ceiling of a positional correlate
In-degree-only cue, no activations	0.500	0.500	0.500	—	identical in the twins (576/576 quartets)

Table 4: Post-hoc variant of the crossing, with a probe trained inside the panel (384 TEST quartets, quartet bootstrap 10,000): canonical marginal AUROC, paired counterfactual ranking $P[s(W^+) > s(W^-)]$ per order, and the interaction (mean paired contrast under permutation minus canonical). The last two rows are logistic models on the records alone, with no activations and no symbol identities, and bound the interpretation: the canonical decoding matches the best a records-only model of position-indexed links can do.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. ProofWriter: Generating Implications, Proofs, and Abductive Statements over Natural Language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3621–3634, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.317. URL https://aclanthology.org/2021.findings-acl.317.

Xiaohan Xu, Chongyang Tao, Tao Shen, Can Xu, Hongbo Xu, Guodong Long, Jian-Guang Lou, and Shuai Ma. Re-Reading Improves Reasoning in Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15549–15575, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.871. URL https://aclanthology.org/2024.emnlp-main.871.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. Do Large Language Models Latently Perform Multi-Hop Reasoning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10210–10229, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.550. URL https://aclanthology.org/2024.acl-long.550.

Sohee Yang, Nora Kassner, Elena Gribovskaya, Sebastian Riedel, and Mor Geva. Do Large Language Models Perform Latent Multi-Hop Reasoning without Exploiting Shortcuts? In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 3971–3992, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.205. URL https://aclanthology.org/2025.findings-acl.205.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations*, 2024.

Honghua Zhang, Liunian Harold Li, Tao Meng, Kai-Wei Chang, and Guy Van den Broeck. On the paradox of learning to reason from data. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 2023.

A Additional tables for the experiments

The tables below support Section 3; the main text reports their decisive values. In the appendix figures, Gemma3-27B and Gemma3-12B denote Gemma-3-27B-it and Gemma-3-12B-it, and S1 denotes the question-mark site.

B Claim-to-evidence map

Table 9 lists every claim with its evidence, its registered status and its scope; secondary and post-hoc results never reclassify a prospective verdict.

Account of the probe	W^-	rewire W^r	frequency-matched W^f	skip W^s
queried truth, type(E,T)	+	0	0	0
wiring of records 3 and 6, position-indexed	+	+	+	0
relation of X to T	+	+	0	0
symbol frequency or prompt length	0	+	0	+
proof depth of type(E,T)		+	+	−

Table 5: Signatures of five accounts on the truth-preserving panel, canonical order: + means W^+ is ranked above the twin, 0 chance, − the reverse. Under permutation, the position-indexed account predicts chance everywhere; the others predict the same row.

Model	Probe	W^-	rewire W^r	frequency-matched W^f	skip W^s	after reordering (four twins)
Qwen3-32B	in-panel	1.000	1.000	1.000	0.990	0.52, 0.53, 0.50, 0.35
Gemma-3-27B-it	in-panel	1.000	1.000	1.000	1.000	0.48, 0.58, 0.53, 0.48
Qwen3-8B	in-panel	1.000	1.000	1.000	0.995	0.46, 0.45, 0.48, 0.37
Gemma-3-12B-it	in-panel	1.000	1.000	1.000	0.992	0.52, 0.57, 0.51, 0.45
Qwen3-32B	pairing (registered)	0.969	0.995	0.971	0.677	0.52, 0.44, 0.52, 0.31
Qwen3-32B	secondary, all cells	1.000	0.802	0.922	0.143	0.54, 0.40, 0.49, 0.27
Gemma-3-27B-it	secondary, all cells	1.000	0.620	0.872	0.445	0.51, 0.45, 0.52, 0.37
Qwen3-8B	secondary, all cells	1.000	0.857	0.943	0.721	0.51, 0.30, 0.48, 0.22
Gemma-3-12B-it	secondary, all cells	1.000	0.648	0.844	0.570	0.44, 0.40, 0.44, 0.41
records only	positional cue	1.000	1.000	1.000	0.500	0.51, 0.51, 0.52, 0.50

Table 6: Probes on the truth-preserving twins, preregistered (384 TEST worlds). Paired ranking $P[s(W^+) > s(\cdot)]$ of the unedited theory over each twin in canonical order, and after the certified reordering in the last column. The in-panel probes separate every twin, including the skip twin, whose records 3 and 6 are those of W^+; this signature matches no account of Table 5 (reading UNCLASSIFIED in all four models). The records-only positional cue has the wiring signature. Gemma-3-27B-it's pairing probe fails the positive control and is omitted. In the two Gemma models the in-panel fit, recorded and refitted alike, stops at the solver's iteration limit (200,000) without meeting its tolerance; the frozen probes are these fits, which reproduce every recorded estimand exactly.

C Complementary preregistered evidence: logical equivalence is not computational equivalence

In the direct-statement and explicit-to-entailed transfer tests, activations are read at the question-mark site of each block ℓ, and a truth direction is estimated from TRAIN pairs of a true and a false case, $u_\ell = \mathbb{E}_b[h^+_{b\ell} - h^-_{b\ell}]/\|\mathbb{E}_b[h^+_{b\ell} - h^-_{b\ell}]\|_2$; projection differences are standardized by the TRAIN standard deviation of paired differences, and blocks are chosen on VALIDATION controls only.

The portability failures above are failures of probes. Four further preregistered experiments ask the behavioral and representational counterpart: does a consequence that is logically implied behave like the same consequence directly stated? Table 11 (Appendix D) summarizes; designs, figures and secondaries are in Appendices D–F.

Direct statement improves access under matching. The direct-statement test (Figure 3) adds the same two symbols in two redundant type records under D0 (target directly stated) and D1 (target one hop away), preserving closure, record count, token length, question-mark-site index, and queried-symbol counts. Both models gain margin from direct statement, in every prespecified order mode; Qwen answers both conditions with literal accuracy 1.00, so the claim concerns access; necessity is not tested. The treatment necessarily binds entity and concept in one record, which we regard as the operational content of direct statement.

Proof distance alone does not predict access. The proof-shortening test reduces a still-unstated target from three hops to one with token-matched redundant records and finds no common graded improvement in either domain, while the direct-statement test shows that crossing to direct statement does: a universal "fewer hops, more access" account is rejected, without establishing a discontinuity.

Model	AUROC	W^-	W^r	W^f	W^s	Reading
Qwen3-32B	0.894	0.951	0.986	0.906	0.863	UNCLASSIFIED
	[0.875, 0.913]	[0.928, 0.971]	[0.973, 0.996]	[0.876, 0.934]	[0.829, 0.896]	
	after reordering	0.557	0.629	0.533	0.538	
Gemma-3-27B-it	0.962	1.000	0.995	1.000	0.475	WIRING ORDER ROBUST
	[0.952, 0.971]	[1.000, 1.000]	[0.987, 1.000]	[1.000, 1.000]	[0.426, 0.525]	
	after reordering	0.559	0.523	0.555	0.462	
Qwen3-8B	0.880	0.987	0.990	0.941	0.685	UNCLASSIFIED
	[0.861, 0.899]	[0.974, 0.997]	[0.979, 0.997]	[0.918, 0.964]	[0.638, 0.732]	
	after reordering	0.479	0.577	0.479	0.596	
Gemma-3-12B-it	0.949	0.999	0.997	1.000	0.973	UNCLASSIFIED
	[0.936, 0.961]	[0.996, 1.000]	[0.992, 1.000]	[1.000, 1.000]	[0.956, 0.987]	
	after reordering	0.523	0.531	0.482	0.576	

Table 7: The model's forced-choice answer on the truth-preserving twins (preregistered; 384 TEST worlds, world bootstrap 10,000). AUROC: canonical marginal AUROC of W^+ against W^- for the mapping-corrected answer gap; paired rankings $P[g(W^+) > g(\cdot)]$ per twin with 95% intervals, in canonical order and after reordering; reading by the rules of Table 5. Gemma-3-27B-it reads WIRING ORDER ROBUST because the one-sided upper bound of its permuted W^- ranking is 0.6003, just above the cut at 0.60. Literal accuracy is near chance through response biases (Qwen3-32B answers "entailed" on 99.8% of W^+ and 94.7% of W^- rows, Gemma-3-27B-it on 9.8% and 0%).

Configuration	Model	Δ gap, canonical [95% CI]	Δ gap, permuted [95% CI]	Over random (canon. / perm.)	Monotone
ProofWriter, frozen direction	Qwen3-32B	$+8.39$ [$+8.15$, $+8.63$]	$+8.29$ [$+8.04$, $+8.56$]	$+9.13$ / $+8.99$	yes / yes
ProofWriter, frozen direction	Gemma-3-27B-it	$+6.51$ [$+5.92$, $+7.08$]	$+6.03$ [$+5.47$, $+6.60$]	$+1.89$ / $+1.32$	no / no
Type-inheritance, frozen probe	Qwen3-32B	-0.11 [-0.16, -0.07]	-0.07 [-0.13, -0.01]	-0.53 / -0.43	decreasing
Type-inheritance, frozen probe	Gemma-3-27B-it	-0.13 [-0.17, -0.09]	-0.07 [-0.12, -0.01]	-0.12 / $+0.37$	no / no

Table 8: Causal steering (directional estimand, post-hoc, fixed before computation). Δ gap: paired change of the mapping-corrected entailment gap between $\alpha = -3$ and $+3$, row bootstrap 10,000 (512 ProofWriter test rows, 384 type-inheritance rows per presentation); "over random": the same change minus that of a random orthogonal direction of equal norm on the same rows. Type-inheritance rows use the IID-selected frozen probe; the second primary probe gives the same pattern.

Transfer exists, but is a small component. Explicit-trained directions transfer to entailed propositions in all four models, yet the pooled entailed effect is an order of magnitude below the explicit controls and depth robustness differs by model. ProofWriter shows robust d1–d3 transfer in both large models but is public and offers no lexical predicate holdout. Full materialization strongly improves behavior while changing length and redundancy; direct target availability is a demonstrated contributor, which is not shown to explain the whole effect.

D THE DIRECT-STATEMENT TEST: DESIGN, RESULTS, AND ROBUSTNESS

Figure 4 shows the manipulation. Each fresh block begins with

$$[\mathrm{type}, e, c_0], \quad [\mathrm{type}, x, c_0], \quad c_0 \to c_1 \to c_2 \to c_3.$$

Both e and x therefore entail both c_2 and c_3 before the experimental records are added. D1 adds $[\mathrm{type}, e, c_2]$ and $[\mathrm{type}, x, c_3]$, leaving the true target $[\mathrm{type}, e, c_3]$ absent but one hop away. D0 swaps the bindings: $[\mathrm{type}, e, c_3]$ and $[\mathrm{type}, x, c_2]$. Every added record is redundant under the base closure.

The experiment contains 64 fresh blocks and 512 forwards/model: two conditions $\times$ true/false query status $\times$ two response mappings. The D0/D1 manipulation concerns the true target; the matched false query remains unreachable in both contexts and acts as a specificity component of the block score. For each condition we average the mapping-corrected correct semantic margin over true/false $\times$ M0/M1. The preregistered per-model success rule requires both $\mathrm{CI}_{95}[\mathrm{score}(D0) - \mathrm{score}(D1)]_{\mathrm{lower}} > 0$ and a positive lower 95% bound for the D0 condition score; the title-level claim is an intersection-union requirement over both large models. The frozen

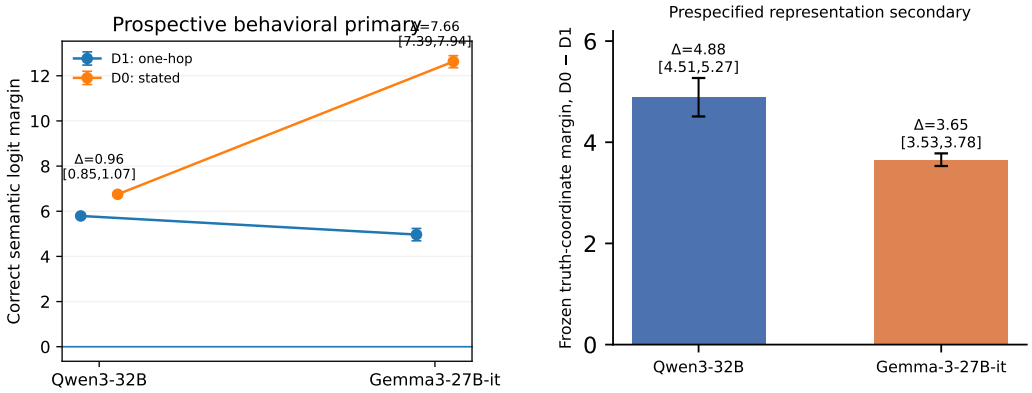

Figure 3: The direct-statement test (Appendix D). **Left:** preregistered behavioral primary; directly stating the target improves the mapping-invariant correct semantic margin in both models. **Right:** prespecified secondary using the truth directions and question-mark-site windows frozen from the explicit-to-entailed transfer test, with no refit or relocalization. Error bars are 95% whole-block bootstrap intervals.

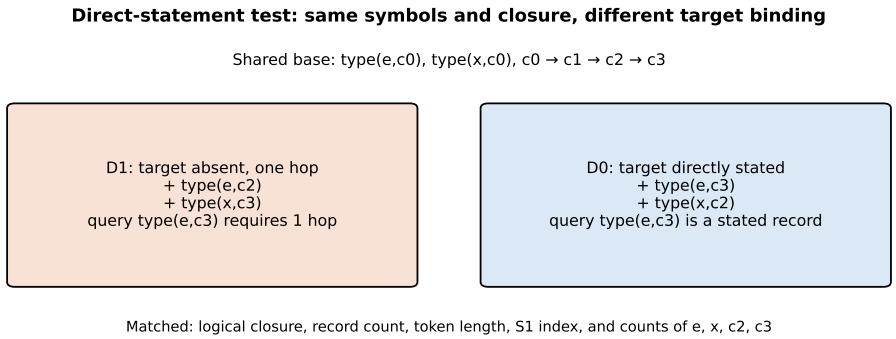

Figure 4: The direct-statement test isolates direct target binding from symbol frequency. D0 and D1 add exactly the same symbols in two redundant type records and preserve the same logical closure. Audits additionally require equal rendered token length, question-mark-site index, queried-symbol counts, and token-footprint lengths for both models. Four order modes balance whether the target entity or target concept stays in the same added-record slot across arms.

explicit-to-entailed transfer test representation contrast D0–D1 is prespecified secondary and cannot rescue the behavioral primary.

The direct-statement test passes the preregistered primary in both models. Qwen3-32B rises from a D1 correct semantic margin of 5.792 (95% CI $[5.693, 5.895]$) to 6.750 ($[6.652, 6.844]$), giving a paired D0–D1 effect of $+0.958$ ($[0.846, 1.072]$). Gemma-3-27B-it rises from 4.968 ($[4.692, 5.238]$) to 12.629 ($[12.356, 12.889]$), for $+7.662$ ($[7.387, 7.944]$). Qwen has literal accuracy 1.00 in both D1 and D0; Gemma rises from 0.855 to 1.00. The behavioral gain is concentrated on the manipulated true proposition: the true-query D0–D1 effect is $+2.022$ ($[1.829, 2.223]$) in Qwen and $+15.260$ ($[14.643, 15.867]$) in Gemma, whereas the false-query effect is -0.106 ($[-0.186, -0.026]$) and $+0.063$ ($[-0.053, 0.179]$). Every prespecified order mode has a positive D0–D1 effect with a positive 95% interval in both models (Qwen: 0.979, 0.869, 0.441, 1.543; Gemma: 8.434, 6.953, 6.512, 8.748 for C-FIXED-HIGH, C-FIXED-LOW, E-FIXED-HIGH, E-FIXED-LOW).

The frozen representation probe shifts even more sharply. In Qwen, the D1 standardized truth-coordinate margin is -0.491 ($[-0.619, -0.355]$) and D0 is 4.395 ($[4.006, 4.809]$), a paired shift of

$+4.885$ ($[4.507, 5.269]$). In Gemma, D1 is 0.002 ($[-0.098, 0.101]$), D0 is 3.657 ($[3.506, 3.803]$), and the shift is $+3.655$ ($[3.531, 3.776]$). These directions were learned on the independent Explicit TRAIN worlds of the explicit-to-entailed transfer test and frozen before the direct-statement test; this is supporting evidence and does not test causal mediation.

E THE PROOF-SHORTENING TEST: DESIGN AND PREREGISTERED OUTCOMES

The proof-shortening test adds exactly one already-entailed type record under three primary conditions on fresh propositions: D3 leaves the target three hops away using a query-irrelevant redundant consequence, D2 adds a target-chain intermediate and leaves two hops, and D1 adds the next intermediate and leaves one hop. D3/D2/D1 have identical logical closure, equal rendered token counts, equal question-mark-site indices, and equal target-symbol token footprints. A fourth D0 cell directly states the target as a preregistered validity control. The preregistered primary fits, separately for representation and behavior, $Y_{bc} = \alpha_b + \beta m_c + \gamma s_c + \epsilon_{bc}$, where $m_c \in \{0, 1, 2\}$ indexes D3/D2/D1 and s_c adjusts the source depth of the added record. The joint claim required a positive lower 95% bootstrap bound for β in both domains and both large models (Table 12); prespecified threshold and endpoint contrasts do not rescue the failed primary.

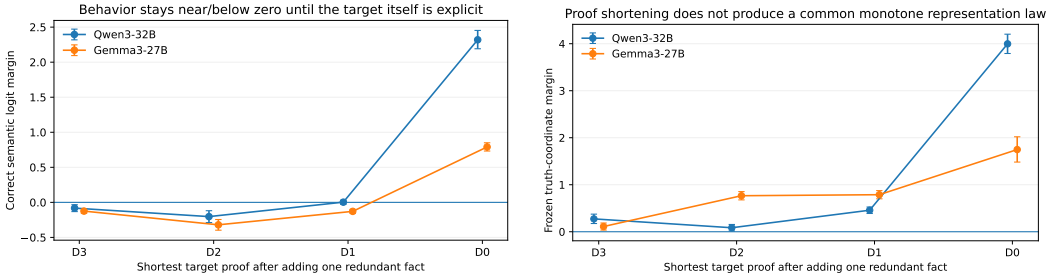

Figure 5: The proof-shortening test: raw condition means. D3/D2/D1 are the preregistered token-matched proof-shortening conditions; D0 directly states the target and served as a validity control outside the primary trend.

F REPRESENTATIONAL TRANSFER AND MATERIALIZATION: DETAILS

In the explicit-to-entailed transfer test, a direction learned only from Explicit TRAIN pairs transfers without refit to pooled Entailed TEST propositions at the question-mark site in both large models: 0.632 ($[0.548, 0.718]$) for Qwen3-32B and 0.364 ($[0.285, 0.454]$) for Gemma-3-27B-it, with full-pipeline permutation $p = .001$ in both. The smaller-model replication prospectively replicates the pooled criterion on fresh worlds in Qwen3-8B, 0.134 ($[0.056, 0.211]$, $p = .006$), and Gemma-3-12B-it, 0.348 ($[0.193, 0.507]$, $p = .001$). Depth decomposition: Qwen3-32B remains positive at d1, d2, and d3 ($0.903, 0.492, 0.502$); Gemma-3-27B-it is positive at d1 (1.352), significantly negative at d2 (-0.409), and inconclusive at d3 (0.149); both smaller models are dominated by d1. Explicit controls are much larger: $8.352, 10.710, 6.029, 7.280$, respectively.

ProofWriter: relearning the direction from ProofWriter Explicit TRAIN and freezing the prospectively selected windows (Qwen 50–52; Gemma 40–42, from Explicit VALIDATION only), the original pooled external gate passes in both large models. A zero-forward decomposition of the stored TEST activations is positive at every tested depth: Qwen 2.295 ($[2.012, 2.574]$), 1.514 ($[1.244, 1.788]$), 1.351 ($[1.142, 1.560]$); Gemma 2.749 ($[2.330, 3.169]$), 1.649 ($[1.202, 2.075]$), 1.372 ($[0.988, 1.756]$) for d1–d3. This decomposition is post-hoc and does not reclassify the prospective verdict of the ProofWriter transfer test. The predicate-overlap parser yields no unseen-predicate stratum (24 normalized predicates in Explicit TRAIN, the same 24 after adding VALIDATION), so lexical predicate holdout is not testable in this sample.

The full-materialization test: a compact theory contains base type facts plus parent rules; the materialized theory retains the same rules and additionally lists every entailed type fact, with identical closure. Across expansion factors 1, 3, 7, and 15, full materialization improves the mapping-invariant

correct semantic margin by $+4.370$ ($[4.162, 4.595]$) in Qwen3-32B and $+3.440$ ($[3.224, 3.665]$) in Gemma-3-27B-it; literal accuracy rises from 78.9% to 100% and from 91.6% to 98.8%. Materialized prompts are 1.6–5.3$\times$ longer and repeat many consequences, so this is not a clean direct-presence intervention (Figure 7).

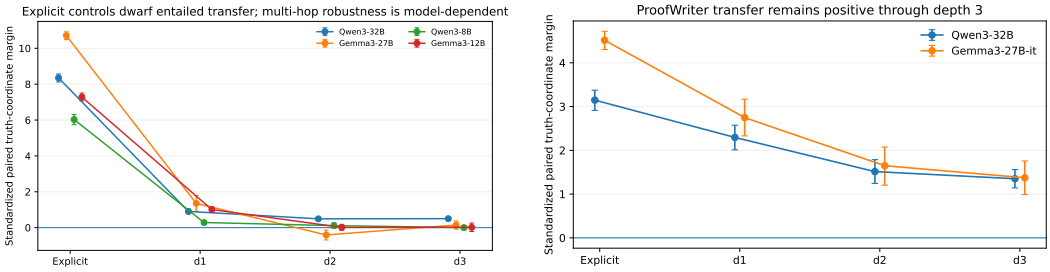

Figure 6: **Left:** Explicit controls and depth-resolved transfer on fresh opaque worlds. **Right:** zero-forward post-hoc depth decomposition of the completed prospective ProofWriter replication with the original frozen directions and windows. Effect magnitudes are standardized within each domain and are not directly comparable across panels.

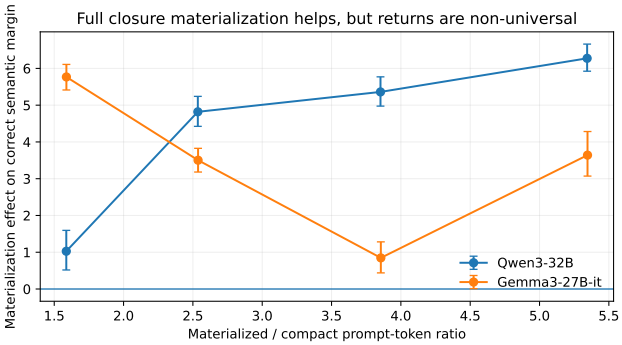

Figure 7: Full closure materialization effect as prompt expansion grows. Error bars are 95% factor-stratified block-bootstrap intervals.

G Model revisions, windows, and forward counts

The direct-statement test uses 512 forwards/model. The proof-shortening test and the full-materialization test use 1,024 forwards/model each. The opaque explicit-to-entailed transfer test and the smaller-model replication use 2,304 forwards/model. The ProofWriter transfer test uses 1,216 forwards/model. The selection-rule development corrected development panel uses 1,664 selection prompts and 192 same-case source-companion prompts per model. Exact code, prompt, activation, and summary hashes are retained in the project research log and provenance files of the supplementary material.

H The fresh-world permutation replication: protocol summary

In this appendix, SOURCE denotes the canonical rendering of the reordering panel and A1 its certified record permutation; R is the label-independent order and R2 a second permutation of it, and TEST-R denotes the TEST cases in label-independent order. P1, P2 and P3 are the preregistered portability, recoverability and label-independent-order outcomes, and S3 is the behavioral secondary (Section 3.1).

Frozen before generation: protocol (hashes in the provenance files of the supplementary material), the selection-rule development base generator module used to build worlds and permutations, the

frozen probes (npz and selector files), and the order-variable diagnostic plan; after generation and token audits, all panels and audits were hashed into a freeze file (the provenance files of the supplementary material) that the forward and analysis scripts verify. Worlds: 576 fresh cases (288 true, 288 false; NEG_PREFIX0/1/2 balanced within every split and truth label), seeds disjoint from earlier panels, eight records each. Panels: C (TEST: canonical source and A1 permutation, identity permutations rejected), R and R2 (all cases: two independent label-independent permutations; TRAIN 288, VALIDATION 96, TEST 192), behavioral (TEST: SOURCE, A1, R; forced-choice question with mappings M0/M1; literal no-space A/B next-token logits). Feature and extraction identical to the selection-rule development (same window indices, statement-period site, 3-block mean). Decision rules: P1 success if the paired case-bootstrap (10,000) 95% upper bound of AUROC(A1)−AUROC(source) < -0.10 for both primary probes, with a source lower bound ≥ 0.80 as positive control (otherwise not evaluable); P2 ESTABLISHED if the one-sided 95% lower bound of the cross-fitted same-family refit ≥ 0.75, with the same refit on the canonical source as positive control; P3 DECODABLE if the TEST-R lower bound of the validation-selected Panel-R probe ≥ 0.85, NOT DECODABLE if its upper bound ≤ 0.65; S3 PRESERVED if literal accuracy under A1 has a lower bound ≥ 0.90, DEGRADED if the paired margin change is entirely below 0. Literal accuracy takes as answer the letter with the larger of the two literal logits; exact ties, which bf16 logits produce in at most 9 of 384 rows per presentation, count as B. Analysis ran once; the Qwen statement-R extraction was interrupted and resumed on the done mask with an identical computation path, recorded in the attempts log.

I THE PROOFWRITER PERMUTATION TEST: PROTOCOL SUMMARY

NO_LOSS, RECOVERED and PRESERVED name the registered readings of three outcomes: the portability of the frozen direction, the direction relearned on permuted theory blocks, and the entailed accuracy after permutation (S3) (Section 3.1).

Frozen before generation: protocol (hashes in the provenance files of the supplementary material), the ProofWriter transfer test audit code that builds prompts, and every artifact of the ProofWriter transfer test used (selected rows, prompts, token audit, summary, prepared rows, activations, logits); panels and token audits hashed into a freeze file (the provenance files of the supplementary material) verified by the forward and analysis scripts. Transformation: full random permutation of the period-terminated sentences of each theory, identity rejected, one permutation per theory block shared by its true and false questions and by both response mappings; certificate: identical sentence multiset, count equal to the number of facts plus rules, whitespace-normalized round trip. Extraction identical to the ProofWriter transfer test (at every decoder block, the question-mark site and the last token of the query line, where the query is known but the question not yet asked, literal A/B logits). Decision rules: positive control (the canonical transfer of the ProofWriter transfer test reproduced: explicit and pooled entailed lower bounds > 0); P1 LOSS if the paired change of the pooled entailed margin has an upper bound < 0, NO_LOSS if its interval includes 0 with a permuted lower bound > 0; P2 RECOVERED if the direction relearned on permuted TRAIN passes the ProofWriter transfer test gate on permuted TEST; S3 PRESERVED if the change in entailed literal accuracy has a lower bound ≥ -0.05 and permuted accuracy ≥ 0.70. Analysis ran once.

J THE WITHIN-WORLD PAIRING CONTROL: PROTOCOL SUMMARY

In this appendix, P0 denotes the registered positive control, P1 and P2 the primary outcomes, and S1 the secondary outcome, which uses a linear SVM on single statements (Section 3.1).

Frozen before forward (protocol and freeze files in the supplementary material): the 576 worlds of the fresh-world permutation replication, splits and seeds; canonical and A1 renderings (A1 regenerated with the frozen base generator and checked equal to the fresh-world permutation replication's A1 views on TEST); two POS statements per world (type(E,T), type(X,T)), exactly one entailed per world by the closure; statement-period extraction identical to the fresh-world permutation replication. Probe: mass-mean direction of the within-world paired difference true minus false at each of the 8 windows, learned on canonical TRAIN (288), window selected on VALIDATION (96) by paired win rate then mean standardized margin, standardized by the TRAIN paired-difference SD. Decision rules: P0 positive control, canonical TEST paired-margin lower bound > 0 and row-level

AUROC one-sided lower bound ≥ 0.80 (failed: 0.63 and 0.51); P1 NO_LOSS / LOSS readings on paired changes A1 minus canonical (not evaluable); P2 direction relearned on A1 TRAIN, RECOVERED if the same bounds hold (NOT_RECOVERED: 0.50 both); S1 secondary, Linear-SVM on single statements, C on validation, canonical 0.93 [0.91, 0.95] and 0.82 [0.77, 0.86], A1 0.52 [0.46, 0.58] and 0.51 [0.45, 0.57]. Registered predictions: shortcut account NO_LOSS; order-dependence account LOSS and NOT_RECOVERED. Analysis ran once.

K THE TRUTH-BY-ORDER CROSSING: PROTOCOL SUMMARY AND REPRODUCIBILITY NOTE

Frozen before forward (protocol and freeze files in the supplementary material): 576 fresh quartets (TRAIN 192, TEST 384), seeds disjoint from earlier panels; W^- is the NEG_PREFIX1 twin of W^+ (records 3 and 6 differ only in their destination symbol); order B is one permutation per quartet shared by W^+ and W^-, identity rejected; the generator checks that record positions and counts of E, X, T, U are identical within each quartet in both orders. Extraction identical to the fresh-world permutation replication. Probes: the within-world pairing control secondary Linear-SVMs (window and C as logged), recomputed from the frozen within-world pairing control activations with `random_state`=0; recomputed validation AUCs 0.8927 and 0.8544 against logged 0.8927 and 0.8545; the recomputed coefficients are stored with their SHA256. Estimands: fraction of quartets with $s(W^+) > s(W^-)$ per order, mean paired contrasts, order shifts per truth, interaction, marginal AUROC per order; quartet bootstrap 10,000. Readings: positive control (canonical marginal AUROC one-sided lower bound ≥ 0.80); ROBUST_PAIRED_DECODING (both rankings' lower bounds ≥ 0.75); PRESENTATION_CONDITIONED (canonical lower bound ≥ 0.75, permuted upper bound ≤ 0.60); TARGETED_SHORTCUT_EVIDENCE (canonical upper bound ≤ 0.60). Secondary refit: same family, C by two-fold CV over TRAIN quartets with all four cells. Reproducibility note (CPU audit, post-hoc). The within-world pairing secondary SVMs were fitted with liblinear without a fixed random seed. Refitting with the original settings and a fixed seed reproduces the stored truth-by-order coefficients exactly (cosine 1.000 in both models). For Qwen3-32B the fit converges, and the dual solver with two seeds and the primal solver give the same coefficients as the stored ones (cosine 1.000; validation AUC 0.893). For Gemma-3-27B-it the original fit does not converge at 20,000 iterations and the dual solver still does not at 200,000; the primal solver converges, and its solution differs from the stored coefficients (cosine 0.935, relative ℓ_2 distance 0.914; validation AUC 0.876 vs. 0.854), yet reproduces the crossing pattern (paired ranking 0.984 canonical vs. 0.513 permuted; marginal AUROC 0.79 and 0.51, so the positive control still fails); the prolonged dual solutions give 0.984 vs. 0.510. The all-cells refit with converged (primal) fits remains not order-robust in both models (1.00 vs. 0.49 and 0.54). None of the 56 Linear-SVM candidates per model in the fresh-world Panel R failed to converge at 20,000 iterations. The coefficients actually used in the crossing are the stored reconstructed ones, computed and hashed before any crossing activation was read; the crossing is prospective for these probes, and its reading does not depend on the convergence of the Gemma fit.

L DEVELOPMENT HISTORY: THE SELECTION-RULE DEVELOPMENT AND THE POST-HOC PERMUTATION DIAGNOSTIC

In this appendix, A1 denotes the certified record permutation and A2 and A3 two other certified operations. The three selection rules compared are the IID rule (validation AUROC), the counterfactual-audit rule (the worst AUROC over the certified operations used for selection) and DIVERSE, its matched control, which selects on an audit set of the same size whose three families vary surface form, context density and graph topology instead of applying certified transformations.

The selection-rule development preregistered a comparison of probe-selection rules (IID, DIVERSE, and counterfactual-audit selection on certified transformations) with a sealed held-out panel. The preregistered primary comparison of the selection-rule development was not evaluated because no candidate passed the original IID-adequacy gate. The cause was a development-panel counterbalancing error in the negative tiers, and the pipeline itself was correct; after correcting it, the unchanged IID gate admitted candidates in both models (40/120 in Qwen3-32B, 56/120 in Gemma-3-27B-it), the prospective comparison was deferred, and the hidden held-out representation panel remained

unopened. A post-hoc development leave-one-operation-out analysis provided no point-estimate evidence that selection on two seen semantic transformations predicted the third better than IID or DIVERSE selection. Everything in this subsection is therefore post-hoc and computed on corrected development worlds; it does not reclassify any prospective verdict.

Same-case permutation collapses frozen probes in both models. For 96 semantic cases (48 true, 48 false) rendered both in the canonical source presentation and under certified record permutation (A1), the frozen IID- and DIVERSE-selected probes fall from ≈0.90 AUROC to 0.54–0.58 in both models (Table 16); paired bootstrap intervals for the A1−source difference exclude zero. The A1 worlds were not intrinsically difficult before permutation.

What the permutation changes in the score. A zero-forward diagnostic (the order-variable diagnostic), with an analysis plan frozen before any computation (Appendix Q), asked whether generator-exposed presentation variables explain the per-case score shift $\Delta s = s_{\mathrm{A1}} - s_{\mathrm{source}}$. Per case, Δz collects order-derived variables of the record list (positions of the query entity's ancestor chain and of the other entity's chain, inversions and topological order, interleaving, disorder relative to the certified source order, positions of records introducing the queried concept), deduplicated by schema rules only; a ridge model $\Delta z \to \Delta s$ is cross-fitted over hash-defined folds with inner regularization selection, and its out-of-fold R^2 is compared with two permutation nulls: a global permutation of Δz across cases and a permutation within truth class. Three results follow. First, the shift is largely class-structured: fold-train class means alone explain 36% of its variance in Qwen3-32B and 15–22% in Gemma-3-27B-it, and, in an exploratory oracle computation that uses the class label, imposing only the cross-fitted class-mean shift on source scores reproduces 77–87% of the observed ranking loss. Second, presentation variables predict the shift no better than class membership does: in Qwen3-32B the out-of-fold R^2 is 0.18 (95% CI $[0.06, 0.27]$), above the global null but at the 95th percentile of the within-class null; in Gemma-3-27B-it it is at most 0.01. Imposing the Δz-predicted shift on source scores reproduces at most 13% of the ranking loss. The supported statement is narrow: the tested presentation variables, with this predictive model and this budget, do not explain the drift beyond the class-conditioned control. It does not exclude a serialization shortcut: the canonical layout of these worlds depends on truth (the record introducing the queried concept and the end of each entity's chain sit at class-specific positions), so a probe may exploit layout information that lies between classes, which the permutation removes and which a within-class control cannot detect; nor does a linear model on position summaries exhaust the functions of order a probe could use. Third, same-family cross-fitted refits on the permuted presentation, over the same 54/42 hash folds (training sets of 42 and 54 cases in 5,000+ dimensions), give AUROC 0.38–0.41 with uncorrelated fold refits (cosine ≈ 0). This establishes the failure of that learning procedure at that budget and does not show that the information is absent: below-chance, fold-inconsistent refits indicate that the fitted directions are dominated by fold-specific structure, for instance the composition of negative tiers. Unlike H2, recoverability is therefore *not established* here; a same-budget refit on the canonical presentation and a generation in which record order is independent of the label from the source presentation on are the tests that settle the reading; Section 3.1 reports them.

M PER-ITEM SCORES AND THE HARMONIZED-PROBE SUPPLEMENT

The supplementary material ships the answer logits of the three forced-choice panels (the fresh-world replication, the truth-preserving twins and the competence panel), with the prompt panels that give their row order, labels and response mappings, and the per-item scores of the direct-statement and materialization tests. Every AUROC, paired ranking and bootstrap interval reported for the model's answer can therefore be recomputed from the supplement. Probe scores are computed from activations, which take several hundred gigabytes and are not shipped; they will be released with the complete code and outputs upon acceptance, and the supplement gives the SHA256 of every activation array. A post-hoc harmonized-probe supplement, computed from the stored ProofWriter activations, trains a Linear-SVM of the family used for the type-inheritance probes on the canonical ProofWriter TRAIN explicit rows, at the frozen question-mark-site window with C selected on validation, and applies it to canonical and permuted TEST rows. Entailed row-level AUROC: 0.969 canonical vs. 0.968 permuted (Qwen3-32B) and 0.939 vs. 0.933 (Gemma-3-27B-it); within-block ranking 0.917 vs. 0.943 and 0.885 vs. 0.891. This is descriptive and was computed after the preregistered analyses; it removes the probe-family difference from the comparison between environments,

leaving the extraction site (question token vs. statement token) and the feature construction (paired questions vs. polarity contrast) as the remaining differences.

N CAUSAL STEERING: PROTOCOL, RUN VALIDITY AND THE REGISTERED ESTIMAND

Directions: for the type-inheritance worlds, the two primary frozen probes of each model (selection development), steered on the frozen forced-choice prompts of the fresh-world panel (384 rows per presentation, canonical and certified permutation); for ProofWriter, the frozen mass-mean truth direction at the question-mark site, steered on the 512 test rows answered with the second response mapping, canonical and permuted. Control: one random unit direction orthogonal to the probe, same σ and layers. Intervention: $\alpha\sigma u$ added to the output of every decoder block of the probe's window, at every prompt position; $\alpha \in \{-3, -1.5, 0, 1.5, 3\}$; answer read from the literal A/B logits at the final position. Validity: unsteered rows must reproduce, row by row, the answer logits recorded by the original forwards. The first Gemma runs failed this check (correlation 0.26–0.44; logits read at the length index under left padding) and were rerun with right padding enforced (correlation > 0.999, answer agreement ≥ 0.99); all Qwen runs passed on the first attempt. Registered estimand: the paired change of the correct-answer margin between $\alpha = -3$ and $+3$, with a criterion requiring a monotone margin, an interval excluding zero and a null random control; no configuration meets it (Qwen ProofWriter $+2.17$ canonical and $+2.33$ permuted with a non-null random control; Gemma ProofWriter -1.74 and -1.84 with a null random control, non-monotone; type-inheritance worlds within ± 0.07). Because a balanced panel makes this estimand cancel for a uniform push toward "entailed", it measures the truth $\times$ steering interaction; the directional estimand of Table 8 was fixed for that reason and logged before it was computed. The post-hoc small-magnitude grid for Gemma on ProofWriter ($\alpha \in \{-1, -0.5, 0, 0.5, 1\}$) reuses the validated plan with only the grid changed; its unsteered rows reproduce the recorded logits exactly (512 of 512 rows in both presentations). The random control's dose–response is non-monotone: both signs of $\alpha = \pm 0.5$ raise the gap above its unsteered value (-8.21 canonically), a response to the perturbation itself.

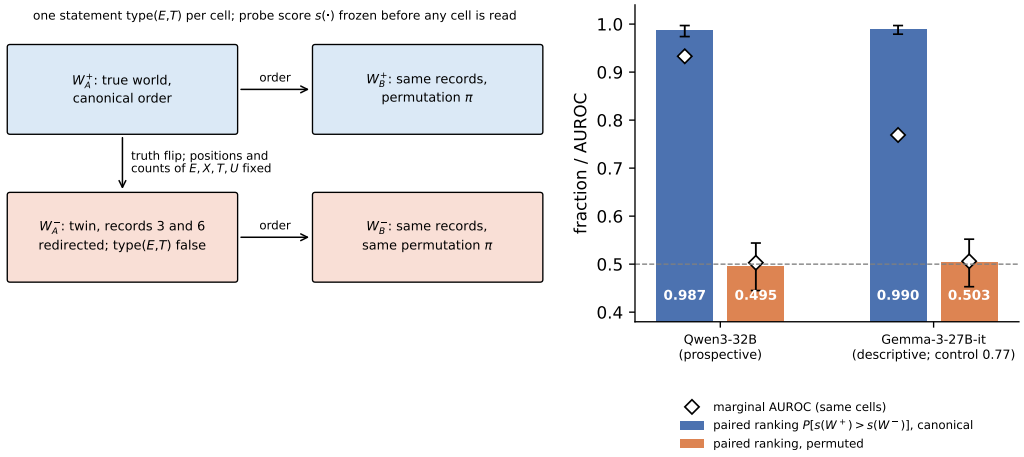

Figure 8: The truth-by-order crossing. Left: a quartet crosses a controlled truth change (the twin redirects records 3 and 6; positions and counts of E, X, T, U fixed) with a record permutation shared by both twins. Right: paired counterfactual ranking of the frozen within-world pairing control probe, $P[s(W^+) > s(W^-)]$ per order with quartet-bootstrap intervals, and the marginal AUROC on the same cells (diamonds); the two measures answer different questions.

O HELD-OUT PRESENTATIONS: ALL SYSTEMS, NEGATIVE TIERS, AND THE SELECTIVITY NULL

Tables 19 and 20 report every system of the held-out presentation test and the per-tier separations that the preregistration requires; both are read from the recorded summary, and no probe was refit. The registered selectivity null was examined by a zero-forward diagnostic run on 17 September on the mixed-unpaired regime of both models: it reproduces the registered null, refits the real selected candidate under TRAIN flips alone, and repeats the full selection with the validation labels flipped as well (Table 18). The mass at AUROC 1 under TRAIN flips alone, together with an almost equal mass at or below 0.10, shows that whole-block flips leave the truth direction in place with a random sign. A synthetic check through the same block-flip pipeline reproduces this when the signal is collinear across blocks (AUROC 1 in 48.7% of draws, at or below 0.10 in 42.1%), and not when there is no signal (null maximum 0.72) or when each block carries its own axis (null maximum 0.82). The diagnostic reclassifies no registered verdict.

P WITHIN-WORLD PAIRING CONTROL: FULL ACCOUNT

The natural explanation of the difference between the fresh-world permutation replication and the ProofWriter permutation test is the training contrast: ProofWriter's true and false questions share one theory, whereas true and false type-inheritance worlds differ in records and layout, so a probe trained across worlds can confound truth with layout. The within-world pairing control registers this as a prediction: in every world of the fresh-world permutation replication exactly one of type(E,T) and type(X,T) is entailed, and both statements are asked about the same theory text, canonical and permuted (2,304 prompts per model). The pairing removes the world-level context difference between classes; it does not remove query-dependent interactions with layout: a logistic predictor on query-relative position metadata (record positions of the queried entity's type record and of the target concept, and the position where the entity's parent chain ends, which is obtained by following the chain and is therefore not a purely superficial cue) reaches 0.89 [0.86, 0.91] AUROC on the within-world labels in canonical order with world-separated folds, and 0.48 under permutation (post-hoc supplement), so the contrast is not free of presentation information. The registered positive control fails in both models: the entity-symmetric mean-difference direction separates the two statements only weakly even in canonical order (row-level AUROC 0.68 Qwen, 0.56 Gemma) and falls to chance under permutation (0.49; relearned on permuted TRAIN, 0.50), so the registered portability test is not evaluable. The preregistered secondary is substantial: a same-family Linear-SVM on the same within-world statements reaches 0.93 [0.91, 0.95] and 0.82 [0.77, 0.86] canonically and 0.52 and 0.51 under permutation. The supported conclusion is that pairing true and false statements within a world does not suffice, with the tested procedures, to produce a portable probe; the crossing of Section 3.2 addresses the positional cues that can be held fixed. Whether portable decodability tracks the model's competence on the property (near-chance forced-choice accuracy here, solved in ProofWriter) was a hypothesis suggested by two configurations; the panel of Section 3.4 tests it prospectively and finds it in three of four models.

Q ZERO-FORWARD POST-HOC DIAGNOSTICS

ProofWriter depth decomposition. It reuses the original width-3 windows selected at the question-mark site of the ProofWriter transfer test, TRAIN Explicit directions, TRAIN standardization, and stored TEST activations. No model is loaded, no window is relocalized, and TEST is not used to refit a direction.

The order-variable diagnostic procedure. The order-variable diagnostic reads only stored activations of the corrected development panel and its source companion; no model is loaded. For each of the 96 same-case pairs and each frozen probe, source and permuted scores are computed from the stored feature contrast and must reproduce the previously recorded source and A1 AUROCs within 0.005 before any inference (reproduction gate). Presentation variables Δz are exported from the generator metadata per case and per statement polarity, then reduced by schema rules alone (missing values, duplicates across polarity, constants, exact aliases, exact linear dependents), leaving 25 columns of full rank. The regression $\Delta z \rightarrow \Delta s$ is a ridge model, standardized on the train-

ing fold, cross-fitted over two folds defined by a salted hash of the world identifier (54/42 cases), with the regularization parameter chosen by inner 5-fold validation on the training fold only. The test statistic is the maximum out-of-fold R^2 over the full model and the group-alone models (positions, relative positions, order, query records, role signatures), recomputed identically in every permutation. Nulls, as frozen in the plan: 2,000 global permutations of Δz and 2,000 permutations within truth class; 5,000 case bootstraps for intervals; control quantile 0.95; $\epsilon = 0.02$ for the reproduction denominator; equivalence margin $\delta_{R^2} = 0.05$; seed 20260919. Reproduction imposes the out-of-fold predicted shift on source scores ($s_{\mathrm{pred}} = s_{\mathrm{source}} + \hat{g}(\Delta z)$) and reports $F_{\mathrm{rep}} = 1 - |\mathrm{AUC}_{\mathrm{pred}} - \mathrm{AUC}_{\mathrm{A1}}|/|\mathrm{AUC}_{\mathrm{source}} - \mathrm{AUC}_{\mathrm{A1}}|$ against three controls (shifts predicted from globally shuffled Δz, from within-class shuffled Δz, and permuted predicted shifts). The secondary refit re-estimates a same-family, same-window, same-C classifier on the permuted presentation of the training fold and scores the held-out fold. Decision rules, thresholds, and the column schema were frozen and hashed before the first computation; verdicts are reported per probe. The exploratory oracle statistic (cross-fitted class-mean shift) was computed afterwards from the saved out-of-fold tables and uses the truth label; it is not part of the frozen plan.

Claim	Evidence	Status	Scope
Changing the presentation (Section 3.1)			
Frozen probe portability across certified H2	The held-out presentation test	Prospective; portability loss 1.000→0.892; refit 0.999	Gemma-3-27B-it; Qwen gate not satisfied
Cross-model method-choice criterion	The held-out presentation test	Preregistered FAIL	separation in Gemma only
Recoverability under permutation at 96 training cases	The fresh-world permutation replication P2	Prospective NOT ESTABLISHED (both models)	canonical positive control ESTABLISHED (0.93 / 0.84)
High linear decodability under label-independent order (this procedure)	The fresh-world permutation replication P3	Prospective NOT DECODABLE AT THIS BUDGET (both models; TEST-R upper bound ≤ 0.65)	validation-selected among 112 candidates, 288 training worlds; TEST-R 0.51 / 0.56
ProofWriter direction under certified sentence permutation	The ProofWriter permutation test P1/P2/S3	Prospective NO LOSS detected, RECOVERED, PRESERVED (both models)	paired change $+0.03$ / $+0.13$, CIs include 0 (not equivalence); cos 0.98 / 0.99
Within-world pairing yields a portable probe	The within-world pairing control (positive control, primaries and secondary)	Prospective: positive control FAILED (mean-difference AUROC 0.68 / 0.56), P1 not evaluable; secondary SVM 0.93 / 0.82 → 0.52 / 0.51	does not suffice with tested procedures; shortcut question open
The answer is truth-sensitive canonically and presentation-conditioned	Behavioral AUROC of the answer gap (post-hoc, zero-forward; plan and gate logged before computation)	Qwen: answer presentation-conditioned, 0.912 → 0.590; Gemma: descriptive, 0.719 → 0.509	registered accuracy near chance through a response bias
The two Qwen primary probes are one probe	Probe agreement (post-hoc, zero-forward; plan and gate logged before computation)	ONE READOUT: Spearman 0.999 in both presentations; Gemma primaries distinct (0.96, 0.90)	Qwen's two primaries count once in P1
Probe and answer share their case-level variation	Probe agreement (post-hoc)	Descriptive: Qwen within-class Spearman -0.01 (true) and 0.18 (false); 0.57 overall in canonical order, -0.04 permuted	IID-selected probes; Gemma 0.19 and 0.45 within class
The H2 loss follows one relation	Per-tier H2 separations, all 18 systems (recorded summary, zero-forward)	SUPPORTED (descriptive): NEG_PREFIX0, the only tier that keeps the parent records, is the least separated in 12 of 14 non-degenerate systems, below both other tiers in 9, inverted in 2	raw score units; compare within a system
The registered selectivity gate tests decodability	Zero-forward diagnostic of the null (post-hoc, 17 September; mixed-unpaired regime)	NOT SUPPORTED: whole-block TRAIN flips randomize the sign of the truth direction; flipping validation labels too still gives p 0.07 to 0.24 at real AUROC 1	exception: Gemma logistic probe, p 0.002 to 0.003
Position metadata predict the within-world labels	The position-metadata supplement (post-hoc)	0.89 [0.86, 0.91] canonical, 0.48 permuted	world-separated folds
Same probe family ports on ProofWriter	Post-hoc supplement (per-item scores)	Descriptive: Linear-SVM trained canonically keeps AUROC and within-block ranking under permutation	0.969→0.968, 0.939→0.933
Behavioral answer under permutation	The fresh-world permutation replication S3	DEGRADED by margin (prospective); literal accuracy near chance in all presentations	format valid (top-1 $\in$ {A,B} 94–100%); margins 1.9→0.35, 1.3→0.17
Changing the truth (Section 3.2)			
The frozen probe separates the crossing's twins in both orders	The truth-by-order crossing (fixed question, truth × order)	NOT SUPPORTED: prospective PRESENTATION_CONDITIONED (Qwen; Gemma descriptive, positive control 0.77)	ranking 0.99 canonical vs. 0.50 permuted; refit with all cells not order-robust
Same profile with an in-panel probe, four models	The crossing, self-contained variant	Post-hoc (designed after the registered crossing; probe frozen before TEST and permuted cells): PRESENTATION_CONDITIONED in Qwen3-32B, Gemma-3-27B-it, Qwen3-8B, Gemma-3-12B-it	canonical 1.000 vs. permuted 0.495–0.542
The probes track the truth of the queried fact	Truth-preserving twins (preregistered)	NOT SUPPORTED: every twin separated in canonical order, truth-preserving ones included (0.990 to 1.000), chance	four models

Claim	Evidence	Status	Scope
From decodability to use (Section 3.3)			
The ProofWriter direction, which passes, is used by the answer (Qwen3-32B)	Causal steering, directional estimand (post-hoc, fixed before computation)	Supported in this configuration only (post-hoc estimand): Qwen $+8.39$ / $+8.29$, monotone, $+9.1$ over random; Gemma $+6.51$ / $+6.03$, above random, non-monotone	registered estimand: criterion not met in any configuration
Gemma steering is specific at a smaller magnitude	Small-magnitude steering, ProofWriter (post-hoc; reading rule logged before computation)	NOT SPECIFIC: probe $+5.66$ / $+5.36$, monotone, $+8.7$ / $+8.2$ over random; random control -3.03 / -2.89, non-null	Gemma-3-27B-it, $\alpha \in \{\pm 0.5, \pm 1\}$
Failing probes are used by the answer	Causal steering, directional estimand	NOT SUPPORTED: -0.07 to -0.25 in both presentations and models, of the order of random	accuracy near chance; the answer's AUROC is 0.91 canonically (Qwen, post-hoc)
The representation the answer uses codes the truth	Interchange interventions at the emergence block (preregistered)	NOT SUPPORTED: REWIRING-CODED in four models, the truth-preserving twin's difference recovered as the false twin's at every block (within 0.06); PRESENTATION-CONDITIONED; probe OFF PATHWAY	last token only; eight dimensions suffice
The registered interchange panel identifies the answer's pathway	Interchange interventions on the twins (preregistered)	NOT INFORMATIVE (registered readings CAUSALLY USED, REWIRING-CODED, PRESENTATION-CONDITIONED and READOUT OFF PATHWAY, met trivially; validity gate G1 amended before any estimate): the selection chose the final token at the last block, where the answer is fixed by construction; readings met trivially	three models; a second registered panel at the emergence layers follows
Competence (Section 3.4)			
Portability follows competence that does not depend on the order	Six cells of depth and negative kind (preregistered)	SUPPORTED in three of four models (COMPETENCE PREDICTS PORTABILITY); VIOLATED BY PORT in Gemma-3-27B-it (three cells)	every competent cell ports; canonically trained probes collapse in every structural cell of every model
Steering along portable probes of competent cells is specific	Second stage of the six-cell panel (preregistered)	NOT SUPPORTED: SPECIFIC in no cell; the six Qwen runs fail the reproduction gate (INVALID), the four Gemma runs read DESCRIPTIVE (random control not null)	ten cells, four models
Development history (Appendix L)			
Counterfactual-audit selection > IID/DIVERSE on a held-out operation	The selection-rule development LOO	No point-estimate evidence	descriptive; no block uncertainty
Record permutation degrades frozen-probe ranking on the same semantic cases	A1 same-case (corrected development)	Post-hoc	both large models; $\approx$0.90$\rightarrow$0.54–0.58
Order/position variables explain the permutation shift beyond class	The order-variable diagnostic (plan frozen before computation)	NOT SUPPORTED	Qwen: class-level only; Gemma: uninformative
Class-level shift accounts for the ranking loss	The order-variable diagnostic oracle	Exploratory (uses the label)	77–87% of loss reproduced
Same-family linear recoverability under permutation	The order-variable diagnostic secondary	Descriptive: not established at 42–54 training cases	AUROC 0.38–0.41; fold refits uncorrelated
Complementary experiments (Appendix C)			
Direct statement > one-hop derivability for behavioral margin	The direct-statement test	Confirmatory PASS	Qwen3-32B, Gemma-3-27B-it
Frozen Explicit truth-coordinate D0>D1	The direct-statement test	Prespecified secondary positive	same two models; not mediation
Common graded D3$\rightarrow$D1 improvement	The proof-shortening test	Confirmatory FAIL	same two models
Pooled Explicit-trained transfer to Entailed	The explicit-to-entailed transfer test	Confirmatory PASS	two large models
Second-scale pooled replication	The smaller-model replication	Confirmatory PASS	Qwen3-8B, Gemma-3-12B-it
Uniform opaque multi-hop transfer	depth-resolved opaque effects	NOT SUPPORTED	only Qwen3-32B clearly positive d1–d3

Experiment	Question (both large models unless noted)	Primary result (95% block-bootstrap CIs)	Verdict
the direct-statement test	Directly stated target vs. one hop away, with identical closure, symbols, token length, and question-mark-site index	Correct semantic margin D0−D1: $+0.958$ $[0.846, 1.072]$ Qwen; $+7.662$ $[7.387, 7.944]$ Gemma. Frozen truth-coordinate shift of the explicit-to-entailed transfer test $+4.885$ / $+3.655$ (secondary).	Confirmatory PASS
the proof-shortening test	Graded benefit of shortening a still-unstated proof (D3→D1), token-matched	Joint positive trend required in both domains and models; Qwen rep. $\beta = 0.033$ $[-0.011, 0.073]$, beh. 0.007 $[-0.017, 0.029]$; Gemma rep. 0.405 $[0.340, 0.475]$, beh. -0.044 $[-0.078, -0.014]$.	Confirmatory FAIL
the explicit-to-entailed transfer test / the smaller-model replication	Explicit-trained truth direction transfers without refit to entailed propositions	Pooled standardized margin 0.632 $[0.548, 0.718]$ Qwen, 0.364 $[0.285, 0.454]$ Gemma ($p = .001$); the smaller-model replication: Qwen3-8B 0.134, Gemma-3-12B-it 0.348. Depth robustness model-dependent.	Confirmatory PASS
the ProofWriter transfer test / depth decomposition	External-domain transfer (ProofWriter)	Pooled gate PASS in both models; d1–d3 all positive in a zero-forward post-hoc decomposition; no unseen-predicate stratum available.	PASS / post-hoc
the full-materialization test	Full closure materialization (same closure, every entailed fact listed)	Margin $+4.370$ $[4.162, 4.595]$ Qwen, $+3.440$ $[3.224, 3.665]$ Gemma; accuracy 78.9→100% and 91.6→98.8%. Not length-matched.	Confirmatory PASS

Table 11: Complementary preregistered experiments. Confirmatory verdicts follow the preregistered rules; secondary and post-hoc results never reclassify them.

Model	Domain	β	95% CI	Primary
Qwen3-32B	representation	0.0330	$[-0.0111, 0.0730]$	FAIL
Qwen3-32B	behavior	0.0069	$[-0.0172, 0.0292]$	FAIL
Gemma-3-27B-it	representation	0.4052	$[0.3395, 0.4749]$	PASS
Gemma-3-27B-it	behavior	-0.0436	$[-0.0777, -0.0144]$	FAIL

Table 12: Preregistered adjusted linear trends across D3/D2/D1. The intersection-union hypothesis required all four cells to pass and therefore failed.

Model	Revision	Question-mark-site window	Role
Qwen/Qwen3-32B	`9216db...c16137`	45–47	large primary
google/gemma-3-27b-it	`005ad3...42228a`	33–35	large primary
Qwen/Qwen3-8B	`b96882...91218`	20–22	scale replication
google/gemma-3-12b-it	`96b6f1...bfd80`	42–48	scale replication

Table 13: Frozen revisions and prospectively selected opaque-world windows. ProofWriter independently selects width-3 windows (Qwen 50–52; Gemma 40–42) from Explicit VALIDATION only. The selection-rule development window indices W0–W7 denote eight three-block groups per model; W05 corresponds to blocks 46–48 in Qwen3-32B.

Model	Selected candidate (of 112)	Validation	TEST-R [95% CI]	TEST-R2	R2−R paired [95% CI]	Reading
Qwen3-32B	W07 Linear-SVM C=0.01	0.513	0.506 [0.425, 0.589]	0.547	$+0.040$ $[-0.074, 0.151]$	NOT DECODABLE at this budget
Gemma-3-27B-it	W06 Logistic-L2 C=10^{-4}	0.553	0.562 [0.481, 0.642]	0.521	-0.041 $[-0.149, 0.068]$	NOT DECODABLE at this budget

Table 14: The fresh-world permutation replication, Panel R (prospective). Candidates trained on 288 worlds under a label-independent record order and selected by validation AUROC (96 worlds); TEST-R: 192 held-out worlds; TEST-R2: a second independent permutation of the same cases; intervals from 10,000 case bootstraps. DECODABLE would require a one-sided TEST-R lower bound ≥ 0.85; NOT DECODABLE requires a one-sided upper bound ≤ 0.65 (observed 0.575 and 0.628). The reading bounds this procedure and leaves other linear directions untested.

Model	Entailed margin, canonical [95% CI]	Entailed margin, permuted [95% CI]	Paired change	AUROC rows (canon. / perm.)	Refit on permuted TRAIN [95% CI]	Accuracy (canon. / perm.)
Qwen3-32B	1.72 [1.57, 1.87]	1.75 [1.60, 1.90]	+0.03 [−0.07, 0.12]	0.975 / 0.971	1.65 [1.50, 1.80]; cos 0.98	0.72 / 0.73
Gemma-3-27B-it	1.92 [1.68, 2.17]	2.06 [1.80, 2.30]	+0.13 [−0.05, 0.32]	0.891 / 0.902	2.51 [2.19, 2.83]; cos 0.99	0.70 / 0.73

Table 15: The ProofWriter permutation test (prospective; ProofWriter, 192 TEST entailed blocks, equal-weight depths d1–d3, 10,000 block bootstraps). Frozen direction and window of the ProofWriter transfer test; readings P1 NO_LOSS, P2 RECOVERED, S3 PRESERVED in both models.

Model	Probe	Source	A1	R^2_{oos} [95% CI]	q_{95} within class	class-means R^2	F_{rep}	oracle shift	A1 refit
Qwen3-32B	IID (W05 SVM)	0.901	0.553	0.18 [0.06, 0.27]	0.19	0.36	0.03	0.87	0.40
Qwen3-32B	DIVERSE (W05 SVM)	0.901	0.553	0.17 [0.06, 0.27]	0.19	0.36	0.03	0.87	0.40
Gemma-3-27B-it	IID (W05 LogReg)	0.900	0.541	0.01 [−0.10, 0.12]	0.08	0.22	0.11	0.82	0.38
Gemma-3-27B-it	DIVERSE (W06 SVM)	0.903	0.576	−0.02 [−0.16, 0.08]	0.03	0.15	0.13	0.77	0.39

Table 16: Same-case permutation diagnostics on corrected development worlds (post-hoc; 96 cases, 48/48). Source and A1: frozen-probe AUROC. R^2_{oos}: out-of-fold R^2 of the cross-fitted ridge $\Delta z \to \Delta s$ (test statistic = max over the full model and group-alone models, recomputed in every permutation); q_{95} within class: 95th percentile of the within-truth-class permutation null. Class-means R^2: fold-train class means alone. F_{rep}: fraction of the ranking loss reproduced by imposing the Δz-predicted shift on source scores. Oracle shift: fraction reproduced by the cross-fitted class-mean shift (exploratory; uses the label). A1 refit: cross-fitted AUROC of a same-family, same-window, same-C refit on the permuted presentation over the 54/42 hash folds (training sets of 42 and 54 cases); it establishes the failure of this procedure at this budget and does not show that the information is absent. Analysis plan frozen before computation (SHA256 `9747be2a...`); no probe was corrected.

Configuration	Model	Order	Δgap probe	Δgap random	false→entailed at α=+3	true→not entailed at α=−3	monotone
ProofWriter, frozen direction	Qwen3-32B	canonical	+8.39 [+8.14, +8.63]	-0.74 [-0.86, -0.63]	18% / 0%	39% / 0%	yes / no
ProofWriter, frozen direction	Qwen3-32B	permuted	+8.29 [+8.03, +8.56]	-0.70 [-0.82, -0.58]	14% / 0%	37% / 0%	yes / no
ProofWriter, frozen direction	Gemma-3-27B-it	canonical	+6.51 [+5.92, +7.08]	+4.61 [+4.39, +4.84]	2% / 74%	14% / 56%	no / no
ProofWriter, frozen direction	Gemma-3-27B-it	permuted	+6.03 [+5.47, +6.60]	+4.71 [+4.49, +4.93]	1% / 66%	17% / 55%	no / no
Type-inheritance, IID probe	Qwen3-32B	canonical	-0.11 [-0.16, -0.07]	+0.42 [+0.38, +0.46]	1% / 4%	0% / 0%	no / yes
Type-inheritance, IID probe	Qwen3-32B	permuted	-0.07 [-0.13, -0.01]	+0.36 [+0.33, +0.40]	0% / 1%	0% / 0%	no / yes
Type-inheritance, IID probe	Gemma-3-27B-it	canonical	-0.13 [-0.17, -0.08]	-0.01 [-0.08, +0.06]	0% / 1%	0% / 0%	no / no
Type-inheritance, IID probe	Gemma-3-27B-it	permuted	-0.07 [-0.12, -0.01]	-0.43 [-0.50, -0.36]	0% / 0%	0% / 0%	no / no

Table 17: Causal steering, specificity of the intervention (post-hoc, from the recorded steering runs; 10,000 row bootstraps). Δgap: change of the mapping-corrected entailment gap between $\alpha = -3$ and $+3$ along the probe and along a random orthogonal direction of equal norm; flip rates (probe / random) are the fractions of false rows answered "entailed" at $\alpha = +3$ and of true rows answered "not entailed" at $\alpha = -3$; monotone: increasing dose–response (probe / random). In Gemma-3-27B-it on ProofWriter the random direction moves the gap by $+4.6$ logits and flips 74% of false rows, so the intervention is not specific at these magnitudes.

		registered null		same candidate, TRAIN flips only			TRAIN and VALIDATION flipped	
Model	Probe	AUROC = 1	p	AUROC = 1	AUROC ≤ 0.10	p	AUROC = 1	p
Qwen3-32B	MassMean	0.467	0.468	0.470	0.476	0.471	0.244	0.245
Qwen3-32B	Logistic-L2	0.269	0.270	0.401	0.437	0.402	0.067	0.068
Gemma-3-27B-it	MassMean	0.448	0.449	0.383	0.445	0.384	0.189	0.190
Gemma-3-27B-it	Logistic-L2	0.002	0.003	0.002	0.060	0.003	0.001	0.002

Table 18: Zero-forward diagnostic of the registered selectivity null (post-hoc, run on 17 September; mixed-unpaired regime, 1,000 permutations each; every real TEST AUROC is 1.000). Share of null draws at AUROC exactly 1 and one-sided p for the registered null (whole-block TRAIN flips, selection on the true validation labels), for the real selected candidate refitted under TRAIN flips alone, and for a null that also flips the validation labels used for selection. For CCS, whose training is unlabeled, the registered null already randomizes the validation orientation (AUROC $= 1$ in 0.279 and 0.143 of draws, p 0.280 and 0.144), and a fixed-window random orientation gives 0.468 and 0.550.

Model	Method	Regime	IID AUROC	IID gate	selectivity p	H1	H2	H3	H4
Qwen3-32B	MassMean	explicit-only	1.000	not passed	0.459	0.600	0.686	0.523	0.905
Qwen3-32B	MassMean	mixed-unpaired	1.000	not passed	0.468	1.000	0.933	0.998	1.000
Qwen3-32B	MassMean	counterfactual-paired	1.000	not passed	0.470	1.000	0.938	0.998	1.000
Qwen3-32B	Logistic-L2	explicit-only	1.000	not passed	0.444	0.599	0.689	0.523	0.908
Qwen3-32B	Logistic-L2	mixed-unpaired	1.000	not passed	0.271	1.000	0.933	0.999	1.000
Qwen3-32B	Logistic-L2	counterfactual-paired	1.000	not passed	0.265	1.000	0.950	0.999	1.000
Qwen3-32B	CCS (linear)	explicit-only	1.000	not passed	0.500	0.639	0.689	0.550	0.896
Qwen3-32B	CCS (linear)	mixed-unpaired	1.000	not passed	0.280	1.000	0.966	0.993	1.000
Qwen3-32B	CCS (linear)	counterfactual-paired	1.000	not passed	0.239	1.000	0.924	0.992	0.986
Gemma-3-27B-it	MassMean	explicit-only	1.000	not passed	0.453	0.713	0.567	0.979	1.000
Gemma-3-27B-it	MassMean	mixed-unpaired	1.000	not passed	0.449	0.901	0.789	0.993	1.000
Gemma-3-27B-it	MassMean	counterfactual-paired	1.000	not passed	0.451	0.886	0.736	0.993	1.000
Gemma-3-27B-it	Logistic-L2	explicit-only	0.999	not passed	0.065	0.662	0.644	0.609	0.792
Gemma-3-27B-it	Logistic-L2	mixed-unpaired	1.000	pass	0.003	1.000	**0.892**	0.945	1.000
Gemma-3-27B-it	Logistic-L2	counterfactual-paired	1.000	pass	0.002	0.993	0.903	0.942	1.000
Gemma-3-27B-it	CCS (linear)	explicit-only	1.000	not passed	0.464	0.722	0.622	0.622	0.939
Gemma-3-27B-it	CCS (linear)	mixed-unpaired	1.000	not passed	0.144	0.516	0.500	0.968	0.729
Gemma-3-27B-it	CCS (linear)	counterfactual-paired	1.000	not passed	0.099	0.516	0.500	0.948	0.688

Table 19: The held-out presentation test, all 18 systems (3 methods $\times$ 3 regimes $\times$ 2 models): IID AUROC, preregistered IID gate, one-sided selectivity p (1,000 refits with the TRAIN labels of whole blocks flipped) and frozen-probe AUROC on the four held-out families; bold marks the one registered L2 failure. The registered cross-model primary (same system passing the gate and failing L2 in both models) was not met. Descriptive table built from the shipped summary; no probe was refit.

System	Family	NEG_PREFIX0	NEG_PREFIX1	NEG_PREFIX2
Gemma-3-27B-it Logistic-L2, mixed-unpaired	H1	2.36 [2.19, 2.54]	3.96 [3.78, 4.13]	3.25 [3.04, 3.45]
Gemma-3-27B-it Logistic-L2, mixed-unpaired	H2	0.36 [0.19, 0.52]	1.86 [1.72, 2.00]	2.00 [1.82, 2.18]
Gemma-3-27B-it Logistic-L2, mixed-unpaired	H3	5.31 [4.34, 6.28]	4.92 [3.97, 5.87]	4.63 [3.70, 5.53]
Gemma-3-27B-it Logistic-L2, mixed-unpaired	H4	5.88 [5.65, 6.08]	7.12 [6.93, 7.30]	5.67 [5.47, 5.86]
Gemma-3-27B-it Logistic-L2, counterfactual-paired	H1	3.71 [3.53, 3.90]	2.61 [2.45, 2.77]	3.99 [3.79, 4.20]
Gemma-3-27B-it Logistic-L2, counterfactual-paired	H2	1.07 [0.88, 1.26]	1.11 [0.97, 1.26]	1.67 [1.48, 1.87]
Gemma-3-27B-it Logistic-L2, counterfactual-paired	H3	5.45 [4.39, 6.47]	5.00 [4.00, 5.98]	4.61 [3.63, 5.56]
Gemma-3-27B-it Logistic-L2, counterfactual-paired	H4	6.21 [5.99, 6.42]	7.12 [6.89, 7.34]	5.65 [5.46, 5.85]

Table 20: TRUE−NEG score difference of the frozen probe (its own units) per negative tier and held-out family (mean over 32 blocks, 95% block-bootstrap intervals), for the systems that passed the IID gate. The H2 loss of the mixed-unpaired Logistic-L2 probe is concentrated in one tier. Reported as the preregistration requires; descriptive.

Model	Probe	Regime	H2 AUROC	NEG_PREFIX0	NEG_PREFIX1	NEG_PREFIX2	L2 failure
Qwen3-32B	MassMean	explicit-only	0.686	0.05 [0.00, 0.09]	0.03 [-0.02, 0.08]	3.82 [3.67, 3.97]	no
Qwen3-32B	MassMean	mixed-unpaired	0.933	2.31 [2.00, 2.62]	10.92 [10.48, 11.36]	12.42 [12.03, 12.81]	no
Qwen3-32B	MassMean	cf-paired	0.938	2.43 [2.11, 2.74]	10.59 [10.15, 11.03]	12.24 [11.86, 12.63]	no
Qwen3-32B	Logistic-L2	explicit-only	0.689	0.00 [0.00, 0.00]	0.00 [-0.00, 0.00]	0.04 [0.04, 0.05]	no
Qwen3-32B	Logistic-L2	mixed-unpaired	0.933	0.13 [0.11, 0.15]	0.61 [0.58, 0.63]	0.80 [0.77, 0.82]	no
Qwen3-32B	Logistic-L2	cf-paired	0.950	0.15 [0.13, 0.17]	0.55 [0.53, 0.57]	0.76 [0.74, 0.78]	no
Qwen3-32B	CCS	explicit-only	0.689	-0.00 [-0.00, 0.00]	0.01 [0.00, 0.01]	0.52 [0.51, 0.53]	no
Qwen3-32B	CCS	mixed-unpaired	0.966	0.00 [0.00, 0.00]	0.01 [0.01, 0.01]	0.01 [0.01, 0.01]	no
Qwen3-32B	CCS	cf-paired	0.924	0.00 [0.00, 0.00]	0.01 [0.01, 0.01]	0.01 [0.01, 0.01]	no
Gemma-3-27B-it	MassMean	explicit-only	0.567	-19.64 [-27.81, -11.52]	6.79 [-0.23, 13.31]	42.69 [33.91, 51.10]	no
Gemma-3-27B-it	MassMean	mixed-unpaired	0.789	15.21 [-25.47, 57.30]	294.08 [216.14, 370.40]	948.54 [888.79, 1007.18]	no
Gemma-3-27B-it	MassMean	cf-paired	0.736	-6.92 [-43.20, 28.78]	161.98 [102.18, 224.09]	819.85 [763.45, 875.67]	no
Gemma-3-27B-it	Logistic-L2	explicit-only	0.644	0.01 [-0.23, 0.24]	0.03 [-0.13, 0.19]	1.26 [1.04, 1.48]	no
Gemma-3-27B-it	Logistic-L2	mixed-unpaired	0.892	0.36 [0.19, 0.52]	1.86 [1.72, 2.00]	2.00 [1.82, 2.18]	yes (gate)
Gemma-3-27B-it	Logistic-L2	cf-paired	0.903	1.07 [0.88, 1.26]	1.11 [0.97, 1.26]	1.67 [1.48, 1.87]	no (gate)
Gemma-3-27B-it	CCS	explicit-only	0.622	-0.00 [-0.01, -0.00]	0.00 [0.00, 0.01]	0.49 [0.48, 0.49]	no
Gemma-3-27B-it	CCS	mixed-unpaired	0.500	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	no
Gemma-3-27B-it	CCS	cf-paired	0.500	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	no

Table 21: TRUE−NEG score difference on H2 per negative tier for all 18 systems of the held-out presentation test (each probe's own units, so tiers are compared within a row; mean over 32 blocks, 95% block-bootstrap intervals; read from the recorded summary). NEG_PREFIX0 negatives swap the two type records and keep the true world's parent records; NEG_PREFIX1 and NEG_PREFIX2 edit parent records. "(gate)" marks the systems that passed the IID gate.