WHEN DOES A NOVEL FORM BECOME A WORD? CAUSAL EVIDENCE FOR REUSABLE LEXICAL FUNCTION IN FROZEN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

When does a novel form begin to function like a word? Inferring its meaning from context is not enough: lexical function emerges when the form itself becomes a cue for reusing semantic information inferred from its prior usages. We study this transition from contextual evidence to lexical function in frozen language models. Across Qwen3-8B and Gemma-3-4B, reversing the meaning assignments of two novel forms reverses which form the model selects for the same held-out situation. We then causally intervene on internal states produced by the preceding usages, separating their influence from the original examples. Transplanted states carry semantic influence from the source context into a new judgment, with a preferential effect on the form with which that meaning was associated. The same states also guide judgments of previously unseen usages, even when the source definition or concept label is no longer presented. This lexical function depends on how the relation was formed, not merely on which semantic material appeared in context. Holding source sentences, order, length, and form frequency fixed, more consistent form–meaning assignments produce stronger transferable semantic influence; subsequent target evidence then reduces the advantage established by stronger formation. Together, these results show that implicit usages do more than make a novel form interpretable: they organize usage-derived semantic information into a relation that can be reused through the form itself. We call this history-dependent, reusable, and evidence-sensitive form–meaning relation a *revisable lexical hypothesis*.

1 INTRODUCTION

What makes an unfamiliar form begin to function like a word? A novel form initially has to be interpreted through its context; after only a few usages, however, language models can learn to use it in previously unseen situations (Eisenschlos et al., 2023; Patel et al., 2023; Wang et al., 2025). For example, after seeing several usages of two novel forms, a frozen language model can be asked to choose which form best fits a new situation. Which form is appropriate depends on the meanings assigned to the two forms by the preceding usages; when those assignments are reversed, the correct choice should reverse as well. Yet even if the model chooses the correct form, this behavior alone does not tell us whether the usages have already produced form-related semantic information that can be reused in a new judgment. The correct answer could instead arise from matching and integrating the original usages only when the model encounters the query.

To determine whether such reusable semantic information has already been formed from the preceding usages, we causally intervene on the internal states produced by those usages and ask whether their influence persists when the original examples are no longer available. We further ask whether this influence is associated with particular forms and whether it generalizes beyond the situations in which those forms were originally observed. This separates two questions: whether context contains information about what a novel form means, and whether that information has acquired *lexical function*—whether it can be reused through the form itself in new semantic judgments.

Across Qwen3-8B and Gemma-3-4B, we find that reversing the meaning assignments of two novel forms reverses which form the model selects for the same held-out situation. More importantly,

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.

internal states produced by the usages carry semantic influence from the source context into new judgments, with a preferential effect on the form with which that meaning was associated. The same states also guide judgments of previously unseen usages, even when the source definition or concept label is no longer presented. The central phenomenon is therefore not merely that context changes what a novel form means. Usage makes the form itself a cue through which semantic information inferred from its prior occurrences can be reused. We call this usage-induced, reusable form–meaning relation a *lexical hypothesis*.

The strength of this lexical relation depends not only on which semantic material appears in context, but on how that material is associated with the form. Holding the source sentences, their order, length, and form frequency fixed, more consistent form–meaning assignments produce stronger transferable semantic influence. The same semantic evidence therefore does not automatically produce the same lexical relation: its causal leverage depends on whether that evidence consistently points to the same form. Nor is this influence fixed once established. As new target evidence becomes directly available, the advantage produced by stronger formation contracts. In this sense, the lexical hypothesis is *revisable*: its influence on later judgments is reweighted as further evidence arrives.

Several lines of prior work establish the ingredients of this problem. Language models can infer unfamiliar meanings from a few in-context examples and generalize them to new usages (Eisenschlos et al., 2023; Wang et al., 2025). Work on internal lexicons has further asked how token sequences come to support word-level representations (Kaplan et al., 2025; Feucht et al., 2024). More broadly, context can induce structured internal representations and bindings, yet forming such representations does not by itself guarantee that they can be flexibly used in later inference (Feng & Steinhardt, 2024; Park et al., 2025; Lepori et al., 2026).

These lines of work leave a question at their intersection: *when does meaning inferred from usage acquire lexical function?* We ask not only whether context can make a novel form interpretable, but whether usage-derived semantic information becomes a reusable causal resource that can be selectively accessed through that form, applied to new situations, and strengthened by a coherent form–meaning history. This connects learning a meaning to using it. A novel form begins to function like a word when it becomes an interface through which semantic information inferred from prior usages can enter new judgments. Frozen language models can therefore acquire temporary lexical function within context without updating their parameters.

2 Research Design and Causal Measurement

Our design separates the formation of a novel form–meaning relation from its later use. We distinguish usage-only source histories, a neutral target carrier receiving the intervention, and held-out semantic judgments. This lets us test whether semantic influence induced by prior usages persists when the original examples are absent, and whether that influence can be selectively accessed through the corresponding form, all without updating model parameters.

2.1 Materials and Experimental Design

Usage contexts with operationally novel forms. We use 300 concepts from CHIME, a dataset of Chinese Internet expressions and their meanings (Xie et al., 2025). Our materials pair each concept with four demonstration usages (indices 0–3) and four held-out usages (q4–q7). In each usage, we replace the original expression with a three-character nonce. The surrounding sentence supplies semantic evidence; the nonce supplies a stable form whose meaning must be inferred from its usages. Exact string and dated-web screening reduce obvious exposure but cannot prove absence from an unknown pretraining corpus, so we use *operationally novel*. Behavioral remapping uses the same q7 situation under both assignments; causal novel-usage tests average over q4–q7. The materials meet the human-review criteria for grammaticality, naturalness, semantic consistency, and leakage (Appendix A.9). A separate English extension using NewTerm evaluates the behavioral and compact-state protocols under independently constructed, human-reviewed materials (Appendix A.10).

Clean reciprocal behavioral test. For every concept pair, the natural grid contains two complete, internally coherent assignments: $M_{AB} : (s_a \to A, s_b \to B)$ and $M_{BA} : (s_a \to B, s_b \to A)$. For a fixed concept, nonce schedule, and held-out q7 situation, only the correct output form reverses.

A strict success requires both assignments. Record-level cross-wiring is analyzed separately as a coherence stress test.

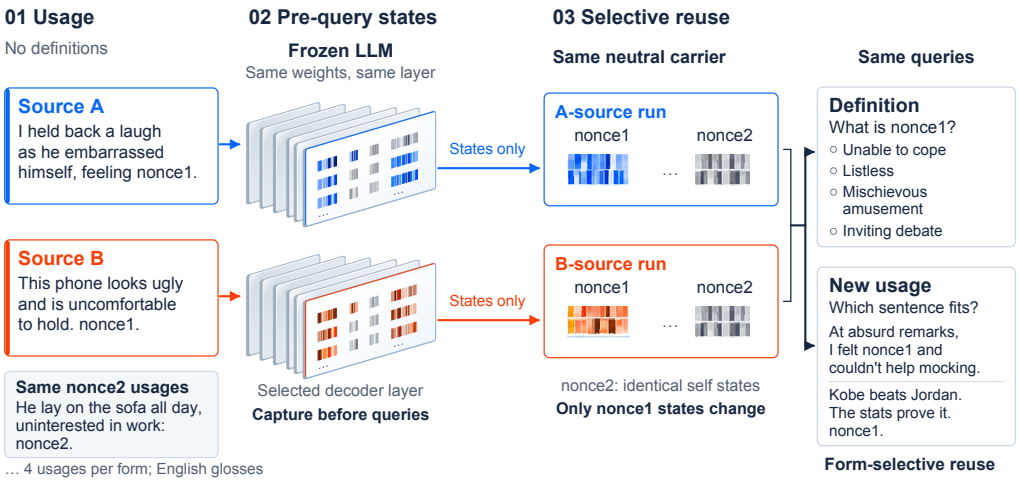

Figure 1: **From usage to form-selective semantic reuse.** Different source histories induce different states for the same novel form. We transplant these usage-induced states into the same neutral carrier while keeping the second form at its exact self state; the original source sentences are absent from the target computation. Definition and held-out novel-usage readouts test whether this semantic influence can be reused and preferentially affects the corresponding form.

Figure 1 illustrates the core causal comparison. Each item contains two novel forms whose usages are interleaved within the same source history. Across self, donor, and random sources, only the concept associated with s_a changes, while the concept associated with s_b remains fixed; the reciprocal orientation reverses their roles. We then transplant the usage-induced states for s_a from either the donor or random source into the same neutral target computation, while keeping s_b at its exact self state. The original source sentences are absent from the target. Let $d \in \{0, 1, 2, 4, 8\}$ denote the number of target usages directly available in the target prompt. At $d = 0$, the neutral carrier still provides four occurrence slots for each novel form, with three nonce tokens per slot; these slots receive the transplanted states but do not themselves provide target semantic evidence. At the intervention interface, each of the four usages contributes the contextualized states of the three nonce tokens, yielding 12 transferred states for the changed form and 12 fixed self states for the unchanged form. This construction lets us ask whether changing the semantic history of one form selectively changes how that form is interpreted while leaving the competing form fixed within the same computation.

We evaluate this transfer in two ways. The definition readout asks which meaning is associated with the queried form, presenting self, donor, random, and fixed-concept meanings in balanced label order. The novel-usage readout instead asks which of two unseen sentences uses the queried form naturally, without displaying the donor definition, concept name, or donor-role label. The latter therefore tests whether the transplanted states support semantic reuse beyond selecting among displayed definitions.

To test whether lexical function depends on how semantic evidence is organized around a form, rather than merely on which semantic material appears in context, we hold the source content fixed. Every source contains the same eight sentences, in the same order, with the same two forms and four occurrences per form. Consistency $c \in \{2, 3, 4\}$ changes only how many usages of each concept are assigned to the corresponding form, with cross-wired slots balanced across reciprocal items. This isolates the contribution of form–meaning consistency while holding semantic content, prompt length, diversity, and form frequency fixed. We then vary d over these registered evidence doses, allowing us to distinguish how strongly the lexical relation is initially established from how much influence that earlier relation retains once new evidence becomes available.

2.2 Intervention and Causal Measures

We formalize the intervention and the causal contrasts used to test whether usage-induced semantic information has acquired lexical function. The core measurements ask whether this information transfers across contexts, acts preferentially through the associated form, generalizes to a different semantic judgment, and varies with formation history and later evidence.

Form-indexed state intervention. For source $X \in \{S, D, R\}$, let $p^X_{ij}(s; c)$ denote token j of the source record aligned to target occurrence slot i of form s by the registered form-preserving record bijection. At the model-specific intervention depth ℓ^*, define the ordered state array

$$H^{\ell^*}_X(s)[i, j] = h^X_{\ell^*, p^X_{ij}(s;c)}, \qquad i = 1, \ldots, 4, \quad j = 1, 2, 3. \tag{1}$$

Each entry is a full-width residual vector, so $H^{\ell^*}_X(s)$ contains the contextualized states from the four observed usages of form s.

For $X \in \{D, R\}$, we construct a hybrid intervention by changing the source of the states associated with s_a while retaining exact self states for s_b:

$$\widetilde{H}^a_X[r, i, j] = \begin{cases} H^{\ell^*}_X(s_a)[i, j], & r = a, \\ H^{\ell^*}_S(s_b)[i, j], & r = b. \end{cases} \tag{2}$$

Let $p^T_{ij}(s_r; d)$ denote the position of constituent token j in occurrence slot i of form s_r in the target sequence at evidence dose d; the superscript T denotes the target. Entry (r, i, j) overwrites the carrier state at this position. All other positions are left unchanged at ℓ^*, after which downstream computation proceeds normally. The alignment preserves form identity and constituent-token order; it does not require absolute source and target positions to coincide. Appendix A.3 gives the complete position operator.

Semantic transfer and form selectivity. For queried form $r \in \{a, b\}$, let $m_r(H; c, d)$ denote the log-probability margin between the donor-consistent and random-control-consistent answers after deploying state H at consistency c and target-evidence dose d. The donor-over-random causal effect is

$$I_r(c, d) = m_r(\widetilde{H}^a_D; c, d) - m_r(\widetilde{H}^a_R; c, d). \tag{3}$$

A positive $I_a(4, 0)$ therefore means that donor states shift the interpretation of the changed form toward the donor meaning relative to matched random states.

Because only the state source for s_a changes, form selectivity is

$$\Delta_{\text{sel}} = I_a(4, 0) - I_b(4, 0). \tag{4}$$

Thus, $\Delta_{\text{sel}} > 0$ asks whether the semantic effect is larger for the changed form than for its unchanged counterpart. This changed-minus-unchanged contrast is our primary selectivity estimand. All effects are measured in log-probability units rather than accuracy points.

Reuse beyond definition selection. To test whether the same transplanted information can support a different semantic judgment, let $m^{\text{ext}}_{a,q}(H)$ denote the log-probability margin between a held-out usage consistent with the donor concept and the matched held-out usage consistent with the random-control concept at query index q, with the queried novel form used in both candidates. We define

$$J_{\text{ext}} = \frac{1}{4} \sum_{q=4}^{7} \left[m^{\text{ext}}_{a,q}(\widetilde{H}^a_D) - m^{\text{ext}}_{a,q}(\widetilde{H}^a_R) \right]. \tag{5}$$

A positive J_{ext} means that the same state intervention shifts judgments of unseen usages toward the donor-consistent use. These outputs contain no donor definition, concept name, or donor-role label. A supporting control under the original multi-depth protocol repeats Equation 5 for the unchanged form and reports the paired changed-minus-unchanged difference (Table 12).

Formation strength and later evidence. With source content fixed, we measure the effect of binding consistency by

$$\Delta_C = I_a(4,0) - I_a(2,0), \qquad \Delta_{CE} = \Delta_C - [I_a(4,8) - I_a(2,8)]\,. \tag{6}$$

Here, $\Delta_C > 0$ asks whether a more consistent form–meaning history produces stronger transferable influence. The interaction $\Delta_{CE} > 0$ asks whether later target evidence reduces the absolute influence advantage established by stronger formation. It therefore measures evidence-dependent reweighting of earlier influence, without identifying a unique internal update mechanism.

The five registered endpoints are

$$\boxed{I_a(4,0) > 0, \quad \Delta_{\text{sel}} > 0, \quad J_{\text{ext}} > 0, \quad \Delta_C > 0, \quad \Delta_{CE} > 0.} \tag{7}$$

Together, they operationalize semantic transfer, form selectivity, cross-judgment reuse, formation strength, and evidence-dependent reweighting. The joint criterion requires positive lower confidence bounds for all five endpoints on the 100 held-out reciprocal-pair clusters in both model families.

2.3 Models, Layer Selection, and Inference

We study Qwen3-8B (Yang et al., 2025) and Gemma-3-4B (Gemma Team, 2025). For the canonical occurrence-state analysis, 50 reciprocal-pair clusters form a discovery split and the remaining 100 form a held-out evaluation split. Layer selection uses only the three discovery endpoints at $c = 4, d = 0$: semantic transfer, form selectivity, and novel-usage reuse. For each endpoint e, we construct a 95% simultaneous max-$|t|$ confidence band across candidate layers and denote its lower bound at layer ℓ by $L_e(\ell)$. A layer is eligible only if all three lower bounds are positive. Among eligible layers, we apply the frozen maximin rule

$$\mathcal{E} = \{\ell : L_e(\ell) > 0,\ \forall e\}, \qquad \ell^* = \arg\max_{\ell \in \mathcal{E}} \min_e \frac{\hat{\mu}_e(\ell)}{\widehat{\text{SE}}_{\text{boot},e}(\ell)}.$$

Ties within 10^{-12} are resolved in favor of the shallower layer. This rule selects layer 15 for Qwen3-8B and layer 11 for Gemma-3-4B. All five registered endpoints are then evaluated on the held-out clusters with 20,000 paired pair-cluster bootstrap draws.

Donor and random controls are fixed within-fold source states rather than synthetic noise. Because an earlier noncanonical run exposed outcomes at the eventually selected layers, we treat the canonical single-layer analysis as locked secondary confirmation; the original multi-depth construct test remains confirmatory. The full evidence hierarchy, source-sharing checks, and sensitivity analyses are reported in the appendix.

3 Selective Semantic Transfer and Reuse

We first establish that novel forms follow the meaning assigned by their current usages, then test whether usage-induced states support form-selective semantic transfer and reuse in new judgments.

3.1 Behavioral Foundation: Contextual Meaning Assignment

Before interpreting internal-state interventions, we first verify that the novel forms behave according to the meanings established by their current usages. The reciprocal test in Section 2.1 reverses the form–meaning assignment while holding the held-out situation fixed.

Both model families follow the current mapping (Table 1). Qwen3-8B reaches 92.00% under M_{AB} and 90.44% under M_{BA}; both directions are correct in 86.67% of matched trials (pair-cluster 95% CI [83.33%, 89.78%]). Gemma-3-4B reaches 72.89% and 71.56%, with strict reciprocal success of 57.33% ([54.33%, 60.44%]). Thus, the form selected for a new situation follows the lexical relation established by the current usages, rather than a fixed preference for the surface form.

Complementary usages strengthen behavioral generalization. On Qwen3-8B, four diverse usages reach 91.22%, compared with 80.17% for exact repetition and 83.17% for paraphrases of a single situation. The corresponding gains are +11.06 pp ([+8.50 pp, +13.67 pp]) and +8.06 pp ([+5.33 pp, +10.78 pp]). These controls match recency, nonce count, and prompt budget.

Table 1: Clean reciprocal q7 remapping over 300 CHIME concepts (150 pair clusters). Strict success requires both mappings; intervals are 95% pair-cluster CIs from 20,000 bootstrap resamples.

Endpoint	Qwen3-8B	Gemma-3-4B
M_{AB} assignment accuracy	92.00% [89.78%, 94.00%]	72.89% [70.11%, 75.67%]
M_{BA} assignment accuracy	90.44% [87.78%, 92.89%]	71.56% [68.78%, 74.33%]
Both mappings correct	**86.67%** [83.33%, 89.78%]	**57.33%** [54.33%, 60.44%]
Pooled clean assignment	91.22% [88.94%, 93.33%]	72.22% [69.83%, 74.61%]
Coherence controls		
Record-level cross-wire reference accuracy	49.22% [46.83%, 51.56%]	50.39% [48.78%, 52.11%]
Clean − cross-wire	**+42.00 pp** [+38.72 pp, +45.17 pp]	**+21.83 pp** [+18.61 pp, +24.94 pp]

Learning depends on a coherent form–meaning assignment. Record-level cross-wiring preserves the semantic materials, order, length, and nonce counts while giving each form conflicting evidence. On Qwen3-8B, accuracy falls from 91.22% in the coherent condition to 49.22% under cross-wiring, a +42.00 pp paired gap (95% CI [+38.72 pp, +45.17 pp]). Gemma-3-4B shows the same pattern: accuracy falls from 72.22% to 50.39%, a +21.83 pp paired gap ([+18.61 pp, +24.94 pp]). Thus, seeing the same semantic material is not sufficient: its consistent assignment to a form substantially improves contextual word use in both model families.

Finally, with all eight sentences, their order, prompt length, and form counts fixed, raising consistency from $c = 2$ to $c = 4$ improves changed-form definition accuracy by 9.33 points on Qwen and 5.11 points on Gemma; held-out extension at $c = 4$ is 87.75% on Qwen3-8B and 72.67% on Gemma-3-4B. Together, these results show that diverse, consistently assigned usages support generalization to new situations. We next use pre-query interventions to examine the semantic content underlying this behavior.

3.2 Form-Selective Semantic Transfer

We now transfer pre-query occurrence states to the neutral carrier, changing one form's source while keeping the other's states exact self. Tests use Qwen layer 15 and Gemma layer 11 on 100 evaluation target/donor pair clusters; selection and inference follow Section 2.3.

Transferred meaning preferentially affects its associated form. Replacing the changed form's nonce-occurrence states at the selected layer shifts definition judgments toward the source meaning. Identity effects are +1.088 ([+0.854, +1.345]) on Qwen and +1.087 ([+0.884, +1.308]) on Gemma. The changed effect exceeds the unchanged effect by +0.572 ([+0.398, +0.759]) and +0.161 ([+0.056, +0.268]), respectively. The corresponding unchanged-form effects are $+0.515$ on Qwen and $+0.926$ on Gemma. Thus both forms are affected, but the changed form receives the larger effect. Selectivity is graded, with a smaller difference between forms on Gemma.

3.3 Reuse in Novel-Usage Judgments

The same single-layer intervention changes which novel held-out usage the model accepts even though neither candidate names the donor concept or exposes a donor role. Pooled q4–q7 extension is +1.455 ([+1.239, +1.680]) on Qwen and +1.573 ([+1.392, +1.763]) on Gemma.

A matched same-interface follow-up additionally queries both forms under the same single-layer intervention. The changed-minus-unchanged novel-usage contrast is +1.592 ([+1.365, +1.827]) on Qwen3-8B and +0.572 ([+0.451, +0.700]) on Gemma-3-4B.

The transferred states therefore carry semantic information usable in a different judgment: choosing a new sentence in which the form is used naturally, with a larger effect on the form whose source

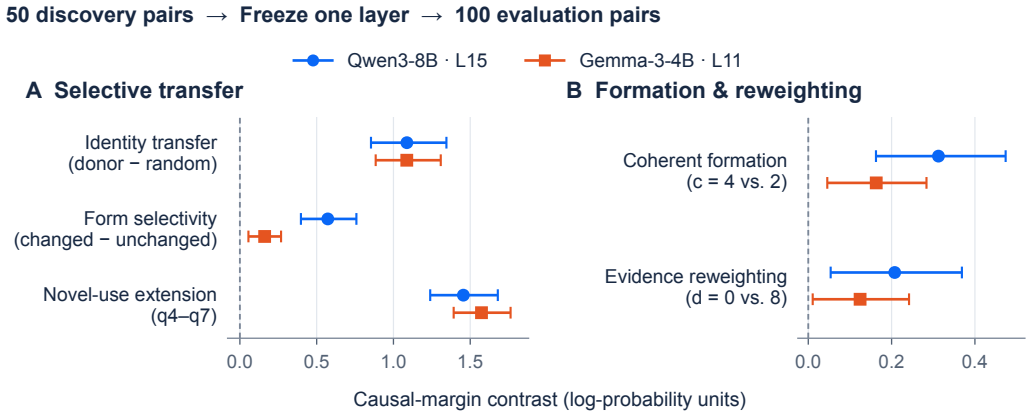

Figure 2: **Pre-query states support selective semantic reuse, formation, and reweighting.** Fifty pair clusters select one layer; 100 disjoint clusters evaluate transfer and reuse (A), formation, and reweighting (B). The hybrid varies 12 changed-form states and retains 12 exact-self states. With sentences fixed, B contrasts four versus two consistent assignments and the contraction of this advantage after target evidence. Bars: 95% CIs, 20,000 paired cluster-bootstrap draws. Selection and inference: Section 2.3.

states were changed. The original multi-depth protocol independently shows the same changed-over-unchanged pattern (Table 12).

Supporting component analyses find the joint signature in nonce-occurrence states but not compact boundary or usage-summary states alone; the original confirmatory multi-depth result and scale extension appear in Appendix B.1– B.2.

4 Usage History and Later Evidence

Having established selective transfer and reuse, we ask what determines the strength of this influence: how form–meaning assignments were established, and what evidence the target receives afterward.

4.1 Binding Consistency and Earlier History

Consistent assignments strengthen the learned binding. With all eight source sentences held fixed, increasing assignment consistency from $c = 2$ to $c = 4$ strengthens the transferred meaning's influence. At zero target evidence, the formation contrasts are +0.313 ([+0.163, +0.474]) on Qwen3-8B and +0.163 ([+0.046, +0.284]) on Gemma-3-4B. The same semantic evidence therefore has greater causal influence when it is consistently associated with the form.

Earlier binding history shapes the last occurrence's causal role. Does this history matter beyond the last local usage? We fix that usage, its nonce token IDs, absolute position, and recent recurrence interval, and change only the form assignments in the earliest two pairs of source usages ($c = 2$ versus $c = 4$). We transfer only the last occurrence's three residual vectors; the other 21 vectors are identical self-$c4$ states in all conditions. Any difference between these conditions must therefore enter through the last occurrence's history-dependent states. The models, selected layers, 100 evaluation pairs, and q4–q7 queries are retained, with no target evidence.

For each queried form r, let $J_{r,\mathrm{last}}(c)$ be the donor-over-random novel-usage effect under this interface. The history contrast is

$$T_{\mathrm{history}} = \underbrace{J_{a,\mathrm{last}}(4) - J_{a,\mathrm{last}}(2)}_{\text{changed-form history effect}} - \underbrace{[J_{b,\mathrm{last}}(4) - J_{b,\mathrm{last}}(2)]}_{\text{unchanged-form history effect}}. \tag{8}$$

Table 2: **Earlier history affects the last occurrence's semantic influence with its local usage fixed.** Three occurrence vectors vary; 21 self-$c4$ vectors are shared. Entries are novel-usage margin contrasts with 95% percentile intervals from 20,000 reciprocal-pair bootstrap draws over 100 pairs per model, pooling orientations and q4–q7 within pairs. The last column is the primary history contrast.

Model	Changed-form history effect	Unchanged-form history effect	T_{history}
Qwen3-8B	+0.12531 [+0.06594, +0.18625]	-0.01375 [-0.05656, +0.03063]	+0.13906 [+0.07250, +0.20469]
Gemma-3-4B	+0.08000 [+0.03047, +0.13047]	+0.05516 [+0.01266, +0.09781]	+0.02484 [-0.02953, +0.07922]

Earlier history increases the changed-form effect in both models (Table 2). Qwen additionally shows a positive selective increment, driven by increased changed-form influence. Gemma shows increases for both forms, but its selective increment is not established. Thus the last occurrence's state conveys an effect of earlier binding history in a new usage judgment, with additional form selectivity supported in Qwen. The effect is measured within the fixed self background. The full protocol and its distinct history schedule are given in Appendix C.1.

4.2 Later Evidence in the Occurrence-State Test

Eight target usages reduce the influence advantage established by consistent assignments: the prespecified Δ_{CE} contrasts are +0.207 ([+0.054, +0.369]) on Qwen and +0.124 ([+0.011, +0.242]) on Gemma.

Common attenuation describes the aggregate consistency surface. Equal-weight rank-one fits of $I_a(c, d) = a(c)g(d)$ to the 15-cell single-layer grids yield RMSE 0.01655 (Qwen) and 0.01411 (Gemma), with positive fitted interactions. Common attenuation thus captures different initial strengths losing influence as evidence grows, without identifying an internal update algorithm. Appendix C.2 reports fitted interactions, pair-level heterogeneity, and multi-depth fits.

4.3 Boundary-Span Extension and Evidence Controls

A separate factorial crosses source usage count $k \in \{1, 2, 3, 4\}$ with evidence dose $d \in \{0, 1, 2, 4, 8\}$. It transplants full-width states from seven fixed token positions appended after the source usages, at seven prespecified depths, into corresponding target boundary positions, without pooling or projection. This boundary-span interface differs from the single-layer, 24-position nonce-occurrence interface in Section 3.2.

The formation advantage contracts with new evidence. At $d = 0$, four usages produce more donor influence than one in both families; later evidence lowers donor influence and raises target margins. The prespecified interaction

$$\Delta_{FE} = [I(4, 0) - I(1, 0)] - [I(4, 8) - I(1, 8)] \tag{9}$$

is positive in both families (Table 3): target evidence reduces the extra influence established by stronger formation.

Later evidence gains influence across models. At four usages, Qwen3-8B, Qwen3.5-4B/9B, and Gemma-3-4B show donor identity, attenuation, and target-evidence gain; Qwen3.5-2B and Phi-4-mini retain only evidence gain (Appendix Table 19). Under the occurrence-state interface, Qwen3.5-2B passes all endpoints except reweighting; 4B and 9B pass all five (Appendix Figure 7).

The evidence response persists at fixed target length. With eight target slots, token length, and intervention positions fixed, both models retain positive target-margin slopes with neutral or irrelevant-natural fillers; neutral fillers also fix nonce frequency (Appendix Table 7). Only Qwen3-8B with neutral fillers retains the full boundary-span identity-and-attenuation pattern: transferred influence is more carrier-sensitive than direct evidence use.

Table 3: Formation and later evidence in the boundary-span interface (300 concepts per model). Each cell gives the estimate [95% CI]. Δ_{FE} is defined in Equation 9; bold estimates pass the stated direction.

Contrast	Qwen3-8B	Gemma-3-4B
Identity $I(4,0) \uparrow$	**+0.5008** [+0.3571, +0.6475]	**+0.0883** [+0.0508, +0.1254]
Formation $I(4,0) - I(1,0) \uparrow$	**+0.3575** [+0.1967, +0.5179]	**+0.0704** [+0.0146, +0.1263]
Attenuation $\beta_{I,4} \downarrow$	**-0.1475** [-0.1909, -0.1044]	**-0.0267** [-0.0391, -0.0140]
Interaction $\Delta_{FE} \uparrow$	**+0.3588** [+0.1917, +0.5288]	**+0.0742** [+0.0121, +0.1367]
Target evidence $\beta_{T,4} \uparrow$	**+8.3231** [+7.9029, +8.7456]	**+3.6727** [+3.4716, +3.8702]

5 Related Work

Learning meanings from context. Prior work studies fast mapping and contextual word learning (Carey & Bartlett, 1978; Renduchintala et al., 2019), novel-word generalization and interpretation (Haley, 2020; Eisenschlos et al., 2023; Patel et al., 2023), and rapid learning through meta-training or generated concept embeddings (Wang et al., 2025; Teehan et al., 2024). Related studies examine lexical representations (Huang et al., 2023; Algayres et al., 2025; Sastre & Rosá, 2026) and context-induced changes in representations and behavior (Park et al., 2025; Lepori et al., 2026; Yona et al., 2026). We test the frozen model's usage-induced states without training an encoder or embedding; unlike prediction-position interventions after a query description (Xu et al., 2026), we capture states from query-free histories.

Word reconstruction and semantic binding. Last-token transplantation recovers whole words, even outside the tokenizer's vocabulary (Kaplan et al., 2025). We instead vary usage-assigned meanings for fixed forms, including after an identical last usage. Entity–attribute binding IDs (Feng & Steinhardt, 2024) and schema rebinding in clone-structured causal graphs (Swaminathan et al., 2023) offer candidate mechanisms for this lexical function.

ICL mechanisms and causal intervention. ICL accounts include function fitting, format cues, induction heads, implicit optimization, and task/function vectors (Garg et al., 2022; Min et al., 2022; Olsson et al., 2022; Akyürek et al., 2023; Hendel et al., 2023; Todd et al., 2024). Pretrained task composition offers another account (Li et al., 2024). We use activation replacement (Vig et al., 2020; Geiger et al., 2021; Meng et al., 2022) to test form-selective meaning reuse while holding the task fixed.

6 Discussion and Conclusion

The lexical-hypothesis account concerns a reusable form–meaning relation, not necessarily a new concept or a localized dictionary entry.

Selective access and scope. Selectivity is graded and weaker on Gemma; wrong-destination effects are mixed (Appendix A.7). Full-state transport separates neither semantic from other contextual content nor identity from semantic codes. Revision changes prefix states' influence, not their contents. The primary test uses Chinese concepts and constrained choices; the human-reviewed English extension supports behavioral binding and boundary-span reweighting under its own protocol (Appendix A.10).

Conclusion. Our results support understanding contextual word learning as the establishment of reusable form–meaning links. Usage-derived states selectively associate semantic information with a particular form, allowing that form to guide judgments in new contexts without parameter updates. A *revisable lexical hypothesis* names this functional link: shaped by usage history, reusable across judgments, and adjustable in its influence as further evidence arrives.

AI USE STATEMENT

Generative AI tools were used to assist with hypothesis development and refinement, provide feedback on experimental methodology, support implementation and debugging, and discuss intermediate results. They were also used to generate experimental text materials, including usage examples and candidate sentences. These materials underwent human review for grammaticality, naturalness, semantic consistency, and leakage. Additional uses included literature retrieval and summarization, manuscript drafting, translation and editing, and scientific figure preparation. The authors reviewed the experimental design, source code, statistical analyses, citations, scientific claims, and final manuscript. Reported numerical results were computed from experimental outputs. The authors take responsibility for the final content of this work, including text, claims, and artifacts produced with AI assistance.

ETHICS STATEMENT

The study analyzes language-model behavior using existing lexical-use data and human-reviewed experimental text materials. It does not solicit sensitive personal information. Human auditing is limited to assessing the grammaticality, naturalness, semantic consistency, and potential leakage of the study materials.

REPRODUCIBILITY STATEMENT

Appendix A specifies the experimental protocols, statistical units, causal interventions, registered endpoints, layer-selection procedure, and integrity checks. The remaining appendices report supporting controls, sensitivity analyses, and the scope of the statistical conclusions.

REFERENCES

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=0g0X4H8yN4I.

Robin Algayres, Charles-Éric Saint-James, Mahi Luthra, Jiayi Shen, Dongyan Lin, Youssef Benchekroun, Rashel Moritz, Juan Pino, and Emmanuel Dupoux. LongTail-Swap: Benchmarking language models' abilities on rare words. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 11231–11251. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.601. URL https://aclanthology.org/2025.findings-emnlp.601/.

Susan Carey and Elsa Bartlett. Acquiring a single new word. *Papers and Reports on Child Language Development*, 15:17–29, 1978.

Hexuan Deng, Wenxiang Jiao, Xuebo Liu, Min Zhang, and Zhaopeng Tu. NewTerm: Benchmarking real-time new terms for large language models with annual updates. In *Advances in Neural Information Processing Systems*, volume 37, pp. 35760–35795, 2024. doi: 10.52202/079017-1128. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/3eec719ab86712d32b065c5977f94ad0-Abstract.html.

Julian Martin Eisenschlos, Jeremy R. Cole, Fangyu Liu, and William W. Cohen. WinoDict: Probing language models for in-context word acquisition. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 94–102, 2023. doi: 10.18653/v1/2023.eacl-main.7. URL https://aclanthology.org/2023.eacl-main.7/.

Jiahai Feng and Jacob Steinhardt. How do language models bind entities in context? In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/9d1b7fc578c0d2d6431fc26d736ecaf3-Abstract-Conference.html.

Sheridan Feucht, David Atkinson, Byron C Wallace, and David Bau. Token erasure as a footprint of implicit vocabulary items in LLMs. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 9727–9739, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.543. URL https://aclanthology.org/2024.emnlp-main.543/.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30583–30598, 2022. doi: 10.52202/068431-2217. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/c529dba08a146ea8d6cf715ae8930cbe-Abstract-Conference.html.

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. URL https://proceedings.neurips.cc/paper/2021/hash/4f5c422f4d49a5a807eda27434231040-Abstract.html.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. URL https://arxiv.org/abs/2503.19786.

Coleman Haley. This is a BERT. now there are several of them. can they generalize to novel words? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 333–341, 2020. doi: 10.18653/v1/2020.blackboxnlp-1.31. URL https://aclanthology.org/2020.blackboxnlp-1.31/.

Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9318–9333, 2023. doi: 10.18653/v1/2023.findings-emnlp.624. URL https://aclanthology.org/2023.findings-emnlp.624/.

Qian Huang, Eric Zelikman, Sarah Li Chen, Yuhuai Wu, Gregory Valiant, and Percy Liang. Lexinvariant language models. *arXiv preprint arXiv:2305.16349*, 2023. URL https://arxiv.org/abs/2305.16349.

Guy Kaplan, Matanel Oren, Yuval Reif, and Roy Schwartz. From tokens to words: On the inner lexicon of LLMs. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/aef75887979ae1287b5deb54a1e3cbda-Abstract-Conference.html.

Michael A. Lepori, Tal Linzen, Ann Yuan, and Katja Filippova. Language models struggle to use representations learned in-context. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14841–14857, 2026. doi: 10.18653/v1/2026.acl-long.676. URL https://aclanthology.org/2026.acl-long.676/.

Jiaoda Li, Yifan Hou, Mrinmaya Sachan, and Ryan Cotterell. What do language models learn in context? the structured task hypothesis. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12365–12379, 2024. doi: 10.18653/v1/2024.acl-long.669. URL https://aclanthology.org/2024.acl-long.669/.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064, 2022. doi: 10.18653/v1/2022.emnlp-main.759. URL https://aclanthology.org/2022.emnlp-main.759/.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022. URL https://arxiv.org/abs/2209.11895.

Core Francisco Park, Andrew Lee, Ekdeep Singh Lubana, Yongyi Yang, Maya Okawa, Kento Nishi, Martin Wattenberg, and Hidenori Tanaka. ICLR: In-context learning of representations. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/83fe5a77502e3d4cfab5960aed0ee6c3-Abstract-Conference.html.

Arkil Patel, Satwik Bhattamishra, Siva Reddy, and Dzmitry Bahdanau. MAGNIFICo: Evaluating the in-context learning ability of large language models to generalize to novel interpretations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2167–2189, 2023. doi: 10.18653/v1/2023.emnlp-main.134. URL https://aclanthology.org/2023.emnlp-main.134/.

Adithya Renduchintala, Philipp Koehn, and Jason Eisner. Spelling-aware construction of macaronic texts for teaching foreign-language vocabulary. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 6438–6443, 2019. doi: 10.18653/v1/D19-1679. URL https://aclanthology.org/D19-1679/.

Ignacio Sastre and Aiala Rosá. Concept tokens: Learning behavioral embeddings through concept definitions. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 26501–26518. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1319. URL https://aclanthology.org/2026.findings-acl.1319/.

Sivaramakrishnan Swaminathan, Antoine Dedieu, Rajkumar Vasudeva Raju, Murray Shanahan, Miguel Lázaro-Gredilla, and Dileep George. Schema-learning and rebinding as mechanisms of in-context learning and emergence. In *Advances in Neural Information Processing Systems*, volume 36, pp. 28785–28804, 2023. doi: 10.52202/075280-1253. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/5bc3356e0fa1753fff7e8d6628e71b22-Abstract-Conference.html.

Ryan Teehan, Brenden M. Lake, and Mengye Ren. CoLLEGe: Concept embedding generation for large language models. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Fkr1yVUb9G.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/4ae163cb8788970e53b4fd9578141139-Abstract-Conference.html.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html.

Wentao Wang, Guangyuan Jiang, Tal Linzen, and Brenden M. Lake. Rapid word learning through meta in-context learning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 32038–32073, 2025. doi: 10.18653/v1/2025.emnlp-main.1631. URL https://aclanthology.org/2025.emnlp-main.1631/.

Yubo Xie, Chenkai Wang, Zongyang Ma, and Fahui Miao. Are large language models chronically online surfers? a dataset for Chinese Internet Meme explanation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 17062–17083, 2025. doi: 10.18653/v1/2025.emnlp-main.863. URL https://aclanthology.org/2025.emnlp-main.863/.

Ningyu Xu, Qi Zhang, Xipeng Qiu, and Xuanjing Huang. Emergent structured representations support flexible in-context inference in large language models. *arXiv preprint arXiv:2602.07794*, 2026. URL https://arxiv.org/abs/2602.07794.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Itay Yona, Amir Sarid, Michael Karasik, and Yossi Gandelsman. In-context representation hijacking. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16852–16867, 2026. doi: 10.18653/v1/2026.acl-long.768. URL https://aclanthology.org/2026.acl-long.768/.

A Experimental Details

A.1 Prompt construction and pairing

Each concept uses fixed target demonstration indices 0–3. Original CHIME expressions are replaced only after usage selection, so selection cannot depend on the sampled nonce. Prompts preserve order across intervention conditions. Queries q4–q7 are independent usage contexts; the evaluated query never contributes to the state or basis used to test it. Same-fold donor concepts are fixed before intervention analysis. For reciprocal behavior, the two concept directions are averaged within each pair before resampling 150 pair clusters.

The state–evidence experiment uses doses 0, 1, 2, 4, and 8. At every dose, state identity changes while prompt structure, intervention positions, tensor shape, and accessible-evidence budget are held fixed. Full-state restoration must reproduce the clean intervened-layer state and margin to absolute tolerance 10^{-6}. Random controls are matched in fold and intervention site.

A.2 Inference and frozen gates

All reported intervals use 20,000 paired bootstrap resamples, with the unit fixed by the design. The formation and compact-state deployment factorials resample 300 concepts. Clean reciprocal behavior and the original multi-depth construct test resample 150 reciprocal-pair clusters. The canonical singleton uses 50 discovery clusters and 100 disjoint held-out pair clusters per model. Here disjointness refers to the main target/donor pairs. Across all source roles, concepts can recur across pairs and splits; the corresponding counts, resampling scope, and all-role deletion analysis appear in Appendix C.3. The follow-up last-occurrence test uses the same 100 evaluation pairs and pair-level aggregation. The arbitration verdict requires: donor minus random state identity at dose zero > 0; its dose slope < 0; and self-target margin dose slope > 0, with all 95% CIs excluding zero. Formation requires a positive coherent log-dose slope, positive four-minus-one contrast, and positive coherent-minus-repeat and coherent-minus-shuffle contrasts. The raw nonce–known moderator test requires a positive nonce-minus-known identity effect, a negative nonce-minus-known dose-slope difference, and a positive known-minus-nonce accuracy difference at zero evidence.

The unified formation–evidence factorial additionally requires five endpoints in each tested family: $I(4,0) > 0$, $I(4,0) - I(1,0) > 0$, a negative evidence-dose slope of $I(4,d)$, a positive $[I(4,0) - I(1,0)] - [I(4,8) - I(1,8)]$, and a positive evidence-dose slope of $T(4,d)$. The specification and its hash were frozen before the 24,000-row full run; no item, layer, position, or endpoint was selected from outcomes.

The lexical construct-validity follow-up freezes the five endpoints in Equation 7. Its inferential unit is the base reciprocal pair: two change orientations and q4–q7 extension readouts are averaged within each of 150 clusters before the paired bootstrap. The complete compact-boundary test was run first and retained after it failed. The distributed form-indexed specification was then frozen as a mechanism-discriminating follow-up before inference. The nonce-only and summary-only decomposition is exploratory and does not alter the endpoint family.

A.3 Causal evaluation procedure

Algorithm 1: evaluation of an inference-time lexical hypothesis. For each target concept, the frozen schedule selects one same-fold donor and one same-fold random concept and rebinds all three source histories to the target surface. First, for each source, formation regime, and dose,

the runner captures the full residual states at $\mathcal{L} \times \mathcal{P}$. Second, it patches donor and random states into the zero-evidence target run to compute $I_\rho(k)$. Third, it crosses every coherent source dose $k \in \{1, 2, 3, 4\}$ with every $d \in \{0, 1, 2, 4, 8\}$ to compute $I(k, d)$ and $T(k, d)$. Fourth, the broader deployment panel patches self, donor, and random coherent-four states into matched target runs. Fifth, it repeats the deployment factorial with the matched original expression. All contrasts are computed within concept; uncertainty uses 20,000 paired concept-level bootstrap resamples. The procedure updates no model parameter and fits no state basis.

Algorithm 2a: original frozen multi-depth interchange. For each reciprocal pair, construct self, donor, and random eight-record histories containing two interleaved forms. Only one form's concept differs; the other four records are identical. Capture seven-depth states at every nonce occurrence and record marker, reorder the eight source blocks to the canonical target form order, and construct a hybrid tensor that takes the four changed blocks from donor/random and the four unchanged blocks from self. Patch the hybrid tensor into an identical neutral carrier. Query both forms with the definition readout; query q4–q7 with the donor-label-free extension readout; then cross $c \in \{2, 3, 4\}$ with $d \in \{0, 1, 2, 4, 8\}$. All target/source block token IDs must match exactly before a task is admitted.

Algorithm 2b: canonical single-layer occurrence interchange. Using 50 discovery pair clusters, test every useful decoder-layer singleton on identity, form selectivity, and pooled q4–q7 extension with simultaneous max-$|t|$ intervals. For each discovery endpoint e and candidate layer ℓ, let $\hat{\mu}_e(\ell)$ denote the estimated effect. We construct endpoint-wise simultaneous 95% confidence bands across layers using

$$T_e^{(b)} = \max_\ell \frac{\left|\hat{\mu}_e^{(b)}(\ell) - \hat{\mu}_e(\ell)\right|}{\widehat{\mathrm{SE}}_{\mathrm{cluster},e}(\ell)},$$

where the cluster standard error in the denominator is held fixed across bootstrap draws. If c_e is the empirical 0.95 quantile of $T_e^{(b)}$, the simultaneous lower bound is

$$L_e(\ell) = \hat{\mu}_e(\ell) - c_e \, \widehat{\mathrm{SE}}_{\mathrm{cluster},e}(\ell).$$

A layer is eligible only when $L_e(\ell) > 0$ for all three discovery endpoints. Among eligible layers, selection uses the maximin score

$$\ell^* = \arg\max_{\ell \in \mathcal{E}} \min_e \frac{\hat{\mu}_e(\ell)}{\widehat{\mathrm{SE}}_{\mathrm{boot},e}(\ell)}, \qquad \mathcal{E} = \{\ell : L_e(\ell) > 0, \ \forall e\},$$

where $\widehat{\mathrm{SE}}_{\mathrm{boot},e}(\ell)$ is the standard deviation of the 20,000 bootstrap estimates. Ties within 10^{-12} are resolved in favor of the shallower layer. Freeze the eligible maximin layer, then evaluate all five construct endpoints on the other 100 clusters. At the selected layer, capture only the three nonce-token states at each of four usages per form. The hybrid contains 24 full-width vectors: the changed form's 12 vectors come from donor or random, while the unchanged form's 12 vectors are exact self. Patch once, allow all downstream computation to proceed normally, and apply the same definition, extension, consistency, and evidence contrasts. No projection, probe, or learned reconstruction is used. Because a previous noncanonical run exposed outcomes at the eventually selected layers, the corrected result is classified as locked secondary confirmation rather than a newly blind confirmatory experiment.

Hybrid reconstruction and block integrity. For the original frozen protocol, we compare a direct exact-self replacement with the algebraic hybrid obtained by supplying self as both variable and fixed block sources. The audit spans all 300 reciprocal units, 11 target prompts, three consistency levels, and seven registered depths. Table 4 reports zero activation, vocabulary-logit, and answer-margin error. A separate static audit performs 21,600 canonical source–target block-ID comparisons per model, checks four changed and four unchanged blocks, bijective record reorderings, and symmetric reciprocal orientations, with no task failure. Source cells are exactly length matched within every causal comparison. Whole source and target sequence lengths need not match because target length changes with evidence and query; transport is restricted to equal-width, token-identical registered blocks, and every donor/random contrast uses the identical target prompt.

Table 4: Hybrid implementation integrity. Each self-reconstruction comparison evaluates a direct exact-self patch and the algebraically reconstructed hybrid on the same target. Errors are maxima over all tasks, targets, consistency levels, vocabulary logits, and registered depths.

Model	Tasks	Trials	Block IDs	Activation	Final logits	Margin
Qwen3-8B	300/300	9900	21600	0.0e+00	0.0e+00	0.0e+00
Gemma-3-4B	300/300	9900	21600	0.0e+00	0.0e+00	0.0e+00

Position-indexed replacement. Let $B_u^X(c)$ be the ordered three-token nonce block of source record u, and $B_v^T(d)$ the corresponding carrier block. The registered permutation $\pi_{X,c}$ maps carrier record v to a source record with the same exact form. It is a bijection over eight records; constituent token j maps to constituent token j, with all three token IDs checked after reordering. For carrier slot (s_r, i) corresponding to record v, $p_{ij}^X(s_r; c) = B^X_{\pi_{X,c}(v)}(c)[j]$ and $p_{ij}^T(s_r; d) = B_v^T(d)[j]$ in Equations 1–2. Equivalently, if Z_T is the unpatched carrier's residual tensor at ℓ^*, the donor or random intervention X is

$$\widetilde{Z}_X[p] = \begin{cases} h^X_{\ell^*, p_{ij}^X(s_a;c)}, & p = p_{ij}^T(s_a; d), \\ h^S_{\ell^*, p_{ij}^S(s_b;c)}, & p = p_{ij}^T(s_b; d), \\ Z_T[p], & \text{otherwise.} \end{cases} \tag{10}$$

Thus the unchanged form receives the same self-source cache in both arms; it is not equated with the carrier's otherwise unpatched activations. Changing c can change the form-to-record assignment and hence the permutation, while changing d can change carrier addresses. The rule and token correspondence are shared across conditions, not the absolute addresses. The released worked example records absolute blocks, token IDs, and the source-to-target permutation for its multi-depth $c = 4$ condition (Appendix A.8); it does not stand in for the full single-layer or E1 registry. E1 instead holds the other 21 vectors at self-$c4$ and changes only the last occurrence, as specified in Appendix C.1.

Table 5: Intervention matrix. Every comparison uses the same frozen pairing schedule and changes only the listed manipulated variable.

Question	Manipulated	Held fixed
Contextual reassignment	Complete reciprocal mapping	Same concept, held-out query, nonce schedule
Evidence coherence	Clean vs. record-level cross-wire	Two forms/concepts, order, frequency, length
Formation	Regime ρ, dose k	Surface, query, interface, token budget
Identity	Self/donor/random source semantics	Surface, shape, depths, positions, query/options
Revision	Target-evidence dose d	Source state, query/options, interface
Fixed-slot semantics	True usages replacing filler	Eight slots, total tokens, absolute positions
Parametric anchor	Nonce vs. known surface	Concept, contexts, source, dose, options
Interface scope	One token vs. 7/8-position span	Semantics, target prompt, depth schedule
Form selectivity	Changed-form blocks donor/random	Unchanged-form blocks exact self
Cross-output reuse	Definition vs. held-out extension	State, carrier, form, pairing
Fixed-content formation	Binding consistency c	Eight sentences, order, length, form counts
Component scope	Nonce vs. marker vs. full blocks	Tasks, layers, endpoints, pairing

A.4 Behavioral controls

Diverse-4 contains four surface-distinct situations expressing one meaning. Exact-repeat-4 repeats one record; Paraphrase-4 varies surface form within one situation; Semantic-shuffle-4 binds fluent situations from different concepts to one nonce. Conditions match order, recency, prompt budget, and nonce occurrences. The locked q7 evaluation uses 150 disjoint reciprocal concept pairs.

For the clean reassignment analysis, each concept–q7–schedule unit is paired across two complete mappings: the target concept is assigned to s_a in one prompt and s_b in the other, while both prompts remain internally coherent. The strict endpoint scores one only when both assignments are answered correctly. Record-level semantic shuffle instead creates conflicting evidence within a form and is therefore treated as a coherence stress test.

Table 6: What each contrast establishes. The joint pattern characterizes a lexical role; it does not uniquely identify the underlying circuit.

Contrast	Supported observation	Interpretive scope
Reciprocal remapping	Answers follow the current form–meaning assignment.	Establishes contextual binding behavior before interpreting state interventions.
Pre-query state transfer	Source activations causally influence the target interpretation.	Semantic information is available before the query; later retrieval of it remains compatible.
Changed minus unchanged form	The changed form receives the larger donor effect.	Establishes relative selectivity, not zero cross-form influence or position-only addressing.
Novel-usage extension	The same intervention changes a second semantic judgment.	Extends beyond displayed-definition selection; outputs remain constrained choices.
Fixed-content consistency	Coherent assignments strengthen transfer with sentences and form counts fixed.	Relates source binding coherence to later causal influence.
Formation-by-evidence interaction	Later evidence narrows the absolute influence gap between consistency levels.	Measures influence-gap reduction, also compatible with common proportional attenuation.

Table 7: Absolute-position-audited fixed-eight-slot control (300 concepts). Bold estimates pass the directional 95% CI gate; all doses are token-length matched, and neutral carriers also fix nonce count. Only one condition passes the complete three-endpoint arbitration gate.

Model	Filler and joint verdict	Portable identity $I(0) \uparrow$	State revision $\beta_I \downarrow$	Target evidence $\beta_T \uparrow$
Qwen3-8B	**neutral nonce (all three)**	**+0.1067** [+0.0363, +0.1762]	**-0.0271** [-0.0505, -0.0041]	**+7.8209** [+7.5079, +8.1242]
Qwen3-8B	natural irrelevant (target only)	+0.0163 [-0.0550, +0.0875]	-0.0053 [-0.0265, +0.0156]	**+7.4947** [+7.1673, +7.8133]
Gemma-3-4B	neutral nonce (target only)	-0.0113 [-0.0488, +0.0258]	+0.0002 [-0.0122, +0.0128]	**+3.5057** [+3.2786, +3.7255]
Gemma-3-4B	natural irrelevant (target only)	-0.0317 [-0.0675, +0.0042]	+0.0040 [-0.0082, +0.0160]	**+3.3056** [+3.1157, +3.4894]

A.5 Fixed-slot semantic-evidence control

Every target prompt contains eight numbered record slots. At dose d, the frozen retained positions contain d target usages and the remaining positions contain filler. Neutral carriers contain the target nonce but explicitly provide no meaning; natural fillers come from the frozen semantic-shuffle registry, excluding the target, donor, and random candidates. All ten cells per item are padded to exactly the same tokenizer length. A separate implementation audit retokenizes every prompt and requires all invariant source and target boundary-token position lists to be identical within concept. The post-alignment audit passes 9,000/9,000 source–target comparisons for each fixed-slot model, with zero minimum, median, and maximum start-position difference. The frozen neutral verdict requires positive donor identity at dose zero, a negative identity slope, and a positive target-margin slope. Natural filler uses the same gates but cannot rescue a failed neutral primary verdict.

For filler type f, write $E_{f,d} = U_T^{(d)} \cup F_f^{(8-d)}$ and $\beta_{T,f} = \mathrm{slope}_{\log_2(d+1)}\, T_f(d)$. With total tokens, absolute positions, and—for neutral carriers—nonce count fixed, $\beta_{T,f} > 0$ identifies target-diagnostic semantics rather than length or surface exposure. It does not by itself establish portable donor identity.

A.6 Notation

A.7 Complete interaction surfaces

Wrong-destination form control. After freezing the aligned analysis, we sent the same four donor/random variable blocks to the unchanged form's occurrence positions while leaving the changed positions exact self. The aligned-minus-misaligned selectivity contrast is mixed across

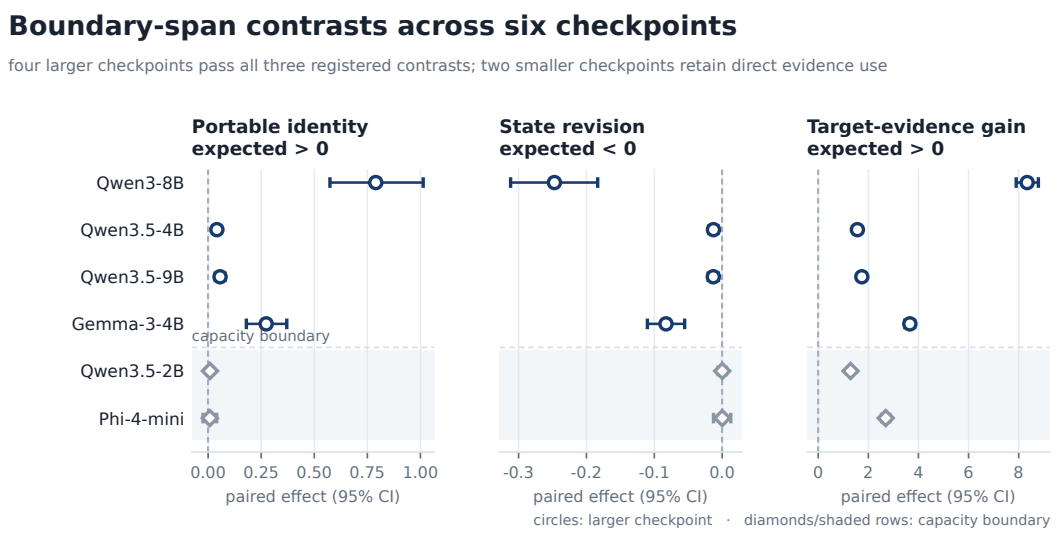

Figure 3: Clean cross-model deployment panel. Circles denote the four larger checkpoints; diamonds and shaded rows denote the two prespecified capacity-boundary checkpoints. Points and bars are concept-level paired effects and 95% bootstrap intervals over 300 concepts. The four larger checkpoints pass portable identity, influence attenuation, and target-evidence gain; the two smaller checkpoints pass only target-evidence gain.

Table 8: Core notation.

Symbol	Meaning
s	Nonce or matched known surface
$U_c^{\rho,k}$	Source usages for concept c, regime ρ, dose k
$S_c^{\rho,k}$	Pre-query residual-state tensor at the tested interface
$H_X(s)$	Distributed source-X states indexed by occurrences of form s
$\mathcal{L}, \mathcal{P}$	Prespecified depths and invariant boundary positions
E_d	Directly accessible target evidence at deployment dose d
$A_\theta(s)$	Parametric lexical constraints associated with s
$I_\rho(k)$	Transferable identity formed from the source history
$I(k,d), T(k,d)$	Donor identity and target margin at formation/evidence doses k, d
$\beta_{I,4}, \beta_{T,4}$	Identity attenuation and target-evidence slopes at $k = 4$
$\Delta_{\text{sel}}, J_{\text{ext}}$	Form selectivity and donor-label-free extension transfer
Δ_C, Δ_{CE}	Fixed-content formation and evidence interaction

readouts and families: it is positive for Gemma's definition readout and Qwen's extension readout, but not for Qwen's definition readout or Gemma's extension readout. Correct destination-position alignment is therefore not a uniformly necessary implementation condition. The main form-selective interchange remains intact, but form identity can travel in transplanted activation content and need not be supplied only by the destination position.

Figure 5 reports all 15 cells of the fixed-content construct-validity surface. Figure 6 reports all 20 cells of the independent coherent-usage-count factorial. Neither conclusion is inferred from endpoint plots alone. In the latter, Qwen's strong–weak gap closes smoothly; Gemma's smaller intermediate cells are not locally monotone, so the text claims the prespecified endpoint interaction rather than monotonic adjacent-dose behavior.

A.8 Worked example

For this item, $I_{\text{changed}} = 7.000 - 4.000 = 3.000$, $I_{\text{unchanged}} = 10.000 - 7.750 = 2.250$, and $\Delta_{\text{sel}} = 0.750$. The donor-label-free q4 extension gives $J_{\text{ext},q4} = 10.750 - 9.000 = 1.750$. With

Table 9: Evidence hierarchy. The complete machine-readable registry records specifications, timestamps, file hashes, estimands, and adjustments. Supporting and exploratory analyses do not upgrade the frozen primary verdicts.

Analysis	Status	Models and unit	Inferential control
Clean reciprocal remapping	Behavioral premise	Qwen3-8B and Gemma-3-4B; 150 pair clusters	Qwen locked derivation; Gemma frozen replication; 20k paired bootstrap
Formation × evidence	Primary confirmatory	Qwen3-8B and Gemma-3-4B; 300 concepts	Five frozen endpoints/model; conjunction; 20k paired bootstrap
Original multi-depth lexical construct	Confirmatory construct validity	Qwen3-8B and Gemma-3-4B; 150 pair clusters	Five frozen endpoints/model; paired cluster bootstrap
Canonical one-layer construct	Locked secondary confirmation	Qwen3-8B and Gemma-3-4B; 50 discovery + 100 held-out clusters/model	Discovery max-$\|t\|$ bands; frozen maximin choice; five held-out endpoints
Qwen3.5 scale extension	Selection-held-out replication	2B, 4B, 9B; 50 discovery + 100 held-out clusters/model	Same singleton rule; 4B/9B pass all five, 2B lacks revision
Compact deployment and fixed-slot tests	Secondary robustness	Six-checkpoint clean panel; two fixed-slot models; 300 concepts	Three-gate conjunction/checkpoint; full fixed-slot gate passes only Qwen neutral
Interface, known-word, and tangent analyses	Supporting or exploratory	Registered model-specific samples	Max-stat/Holm only for their stated secondary families; no rescue of primary claims
NewTerm-English	Supporting external extension	Qwen3-8B and Gemma-3-4B; 300 concepts	Human-reviewed materials; behavior and compact-state protocol only

Table 10: Selective interchange by reciprocal orientation. Intervals are raw 95% paired cluster-bootstrap CIs over 150 base pairs.

Model	Changed binding	I_{changed}	$I_{\text{unchanged}}$	Δ_{sel}
Qwen3-8B	s_a changed; s_b self	+2.955 [+2.522, +3.403]	+2.298 [+1.953, +2.670]	+0.657 [+0.430, +0.887]
Qwen3-8B	s_b changed; s_a self	+3.155 [+2.707, +3.630]	+2.493 [+2.115, +2.880]	+0.662 [+0.397, +0.928]
Gemma-3-4B	s_a changed; s_b self	+2.012 [+1.637, +2.428]	+1.459 [+1.123, +1.828]	+0.553 [+0.347, +0.770]
Gemma-3-4B	s_b changed; s_a self	+2.267 [+1.856, +2.712]	+1.702 [+1.371, +2.057]	+0.565 [+0.338, +0.810]

content fixed, $\Delta_C = 1.000$ and $\Delta_{CE} = 1.625$. These item-level values illustrate the estimands; all conclusions use the 150-cluster paired analysis.

The supplement releases a frozen worked example of the original multi-depth construct test as `supplement/worked_example/worked_example_full.json` and an UTF-8 browser rendering. These artifacts include the complete Chinese self, donor, and random histories; neutral carrier; changed/unchanged definition and q4–q7 extension prompts; every option and label permutation; full tokenizer IDs and tokens; the eight absolute position blocks; seven relative depths; block provenance; raw role log probabilities; and explicit calculations of m_r, $I_r(c, d)$, Δ_{sel}, J_{ext}, Δ_C, and Δ_{CE}. Hashes of every source result shard and generated artifact are recorded in the adjacent manifest.

A.9 Human review of study materials

The primary audit covers four demonstrations and q4–q7 for each of 300 concepts, evaluating grammaticality, naturalness, semantic consistency, and leakage. Human review of the study materials

Table 11: Supporting cross-form addressing control. Misaligned insertion sends the same donor/random variable blocks to the unchanged form's occurrence positions while the changed positions remain exact self. Intervals use 20,000 paired bootstrap resamples over 150 reciprocal-pair clusters.

Readout	Model	Aligned selectivity	Misaligned selectivity	Aligned − misaligned
Definition	Qwen3-8B	+0.659 [+0.459, +0.869]	+0.837 [+0.692, +0.986]	-0.177 [-0.400, +0.037]
Definition	Gemma-3-4B	+0.559 [+0.395, +0.729]	+0.120 [+0.007, +0.239]	+0.439 [+0.278, +0.604]
Extension	Qwen3-8B	+2.203 [+1.941, +2.480]	+0.851 [+0.727, +0.977]	+1.352 [+1.127, +1.587]
Extension	Gemma-3-4B	+0.853 [+0.710, +0.997]	+1.066 [+0.927, +1.207]	-0.213 [-0.378, -0.051]

Table 12: Held-out extension effects by query under the original multi-depth protocol (150 reciprocal-pair clusters). Raw intervals are paired cluster-bootstrap CIs; the final column uses a simultaneous max-stat 95% interval across q4–q7. Holm-adjusted one-sided p-values apply to changed-minus-unchanged selectivity.

Model	Query	J_{changed} raw	$J_{\text{unchanged}}$ raw	Difference raw	p_{Holm}	Difference simultaneous
Qwen3-8B	q4	+2.828 [+2.537, +3.135]	+0.642 [+0.486, +0.805]	+2.186 [+1.846, +2.532]	0.0002	+2.186 [+1.747, +2.625]
Qwen3-8B	q5	+3.075 [+2.763, +3.392]	+0.810 [+0.632, +0.992]	+2.265 [+1.924, +2.619]	0.0002	+2.265 [+1.826, +2.704]
Qwen3-8B	q6	+2.812 [+2.484, +3.152]	+0.669 [+0.478, +0.854]	+2.142 [+1.792, +2.497]	0.0002	+2.142 [+1.703, +2.582]
Qwen3-8B	q7	+2.974 [+2.632, +3.332]	+0.755 [+0.584, +0.929]	+2.219 [+1.845, +2.605]	0.0002	+2.219 [+1.780, +2.658]
Qwen3-8B	pooled	+2.922 [+2.674, +3.179]	+0.719 [+0.604, +0.836]	+2.203 [+1.937, +2.484]	–	–
Gemma-3-4B	q4	+2.810 [+2.482, +3.153]	+2.120 [+1.819, +2.423]	+0.690 [+0.414, +0.968]	0.0002	+0.690 [+0.338, +1.042]
Gemma-3-4B	q5	+2.809 [+2.476, +3.152]	+1.913 [+1.625, +2.202]	+0.896 [+0.637, +1.160]	0.0002	+0.896 [+0.543, +1.248]
Gemma-3-4B	q6	+2.998 [+2.639, +3.370]	+2.087 [+1.759, +2.432]	+0.910 [+0.625, +1.196]	0.0002	+0.910 [+0.558, +1.263]
Gemma-3-4B	q7	+3.032 [+2.682, +3.405]	+2.115 [+1.812, +2.430]	+0.917 [+0.635, +1.200]	0.0002	+0.917 [+0.565, +1.270]
Gemma-3-4B	pooled	+2.912 [+2.674, +3.158]	+2.059 [+1.859, +2.266]	+0.853 [+0.709, +0.997]	–	–

is complete and the materials meet the stated requirements. The reported experiments retain their original demonstration and query indices; this revision adds no q8/q9 or paraphrase experiment.

A.10 English external extension

To test scope without changing the primary estimand, we applied the frozen behavior and clean arbitration designs to 300 concepts from NewTerm (Deng et al., 2024) (our NewTerm-English extension). An independent generator produced d0–d3 and q4–q7 from source meanings; every nonce has exactly three tokens under all evaluated Qwen and Gemma tokenizers. The completed machine grid meets its numerical gates and human review of the materials is complete. Table 16 reports the existing behavior and compact-state extension. It assesses cross-language scope for those protocols, rather than replicating the full two-form singleton test.

Both Qwen3-8B and Gemma-3-4B pass the unchanged English behavior and causal gates. In a separate cross-fitted surface comparison, states are formed only from d0–d3 while q4/q5 no-binding behavior measures lexical anchoring. The paired nonce-minus-original identity effect is positive on both models, so the surface-status contrast also holds outside CHIME. The stronger scalar claim does not: the pooled continuous anchor coefficient crosses zero. Thus the extension is consistent with original lexical constraints moderating causal reliance, but does not establish a linear law from measured no-binding strength to state controllability.

B Additional Mechanistic Results

B.1 Canonical singleton and Qwen3.5 scale extension

The main analysis tests every useful single layer on 50 discovery reciprocal-pair clusters, freezes the eligible maximin layer, and evaluates all five construct endpoints on 100 disjoint clusters. The canonical interface is therefore selected by a fixed causal criterion rather than chosen from the held-out curve. It contains only the three contextualized tokens at each of the eight nonce occurrences: 24 full-width residual vectors in total, of which 12 belong to the changed form and 12 remain exact self states.

Table 13: Complete fixed-content consistency-by-evidence surface $I(c, d)$ under the original multi-depth protocol; 95% paired cluster-bootstrap CIs. No local monotonicity claim is made.

Model	Consistency	$d = 0$	$d = 1$	$d = 2$	$d = 4$	$d = 8$
Qwen3-8B	$c = 2$	+2.547 [+2.238, +2.880]	+1.514 [+1.220, +1.847]	+1.046 [+0.788, +1.342]	+0.616 [+0.430, +0.827]	+0.425 [+0.288, +0.576]
Qwen3-8B	$c = 3$	+2.772 [+2.423, +3.136]	+1.645 [+1.320, +2.001]	+1.088 [+0.823, +1.391]	+0.640 [+0.453, +0.843]	+0.507 [+0.368, +0.665]
Qwen3-8B	$c = 4$	+3.055 [+2.672, +3.471]	+1.896 [+1.550, +2.284]	+1.254 [+0.960, +1.589]	+0.768 [+0.583, +0.973]	+0.586 [+0.425, +0.773]
Gemma-3-4B	$c = 2$	+1.885 [+1.624, +2.177]	+1.231 [+0.993, +1.489]	+1.058 [+0.851, +1.278]	+0.742 [+0.569, +0.933]	+0.497 [+0.366, +0.644]
Gemma-3-4B	$c = 3$	+2.066 [+1.759, +2.402]	+1.342 [+1.081, +1.630]	+1.227 [+0.992, +1.487]	+0.844 [+0.648, +1.065]	+0.539 [+0.404, +0.693]
Gemma-3-4B	$c = 4$	+2.140 [+1.822, +2.484]	+1.443 [+1.167, +1.750]	+1.309 [+1.051, +1.597]	+0.890 [+0.681, +1.123]	+0.552 [+0.417, +0.702]

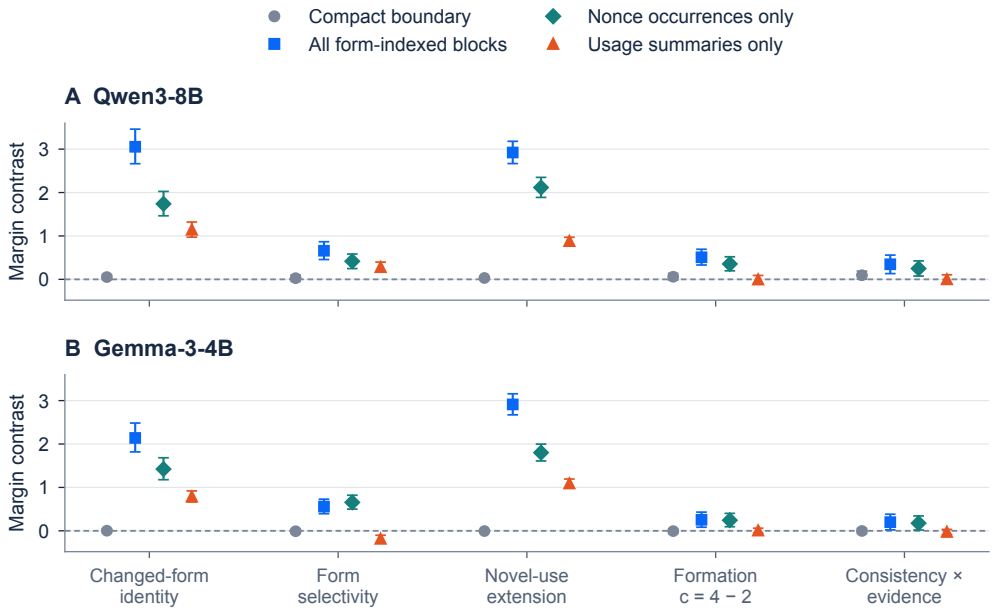

Figure 4: Component decomposition. Points show paired causal contrasts with 95% cluster-bootstrap CIs. Only the full form-indexed blocks and distributed nonce-occurrence states pass all five endpoints in both models. This is an exploratory implementation decomposition; the endpoint family was unchanged.

We apply the identical selection and held-out protocol to Qwen3.5-2B, -4B, and -9B. The rule selects layers 7, 11, and 11, respectively. All three sizes pass the held-out identity, form-selectivity, and label-free extension gates. The 4B and 9B models also pass formation and revision; the 2B model passes formation but its revision interval crosses zero. This capacity boundary does not alter the two-family main result.

B.2 Original multi-depth and layer-budget corroboration

The original frozen multi-depth test passes all five construct endpoints on all 150 reciprocal-pair clusters in both model families (Table 18). We retain it as confirmatory corroboration while using the more economical, discovery-selected singleton as the canonical main-text interface.

We preregistered an implementation audit of the nonce-occurrence interface before inspecting any new layer result. It retains the same selective hybrid and three existing transport endpoints—changed-form identity, form selectivity, and pooled q4–q7 extension—while sweeping every decoder-layer singleton, fixed architecture-relative budgets $k \in \{1, 2, 3, 4, 5, 7, 9, 12\}$, five seven-useful-depth controls, and all useful layers. Fifty reciprocal-pair clusters formed the discovery split and 100 disjoint clusters the confirmatory split in each model; all intervals use 20,000 paired cluster-bootstrap draws. Because a pre-query occurrence replaced after the final attention block cannot affect the query row, the registered legacy schedules are nominally seven but effectively six interventions. The final-layer singletons

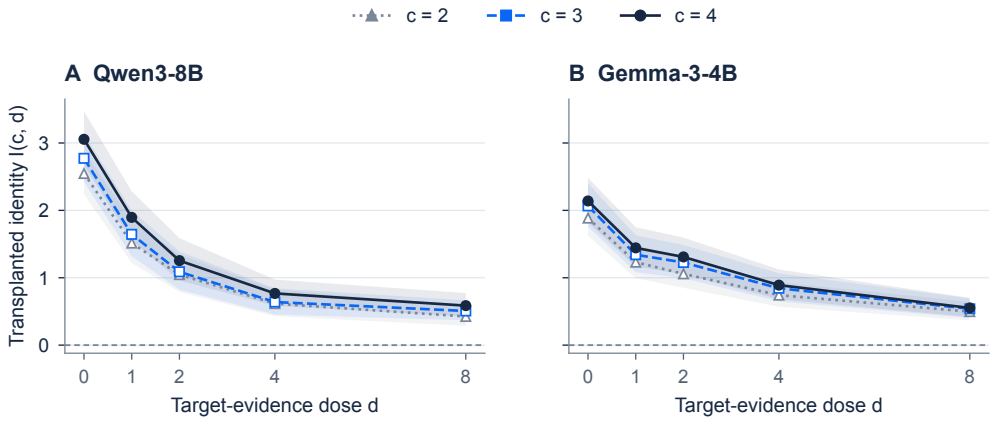

Figure 5: Complete fixed-content consistency-by-evidence surfaces under the original multi-depth protocol. Lines are $I_a(c, d)$ with 95% paired cluster-bootstrap CIs.

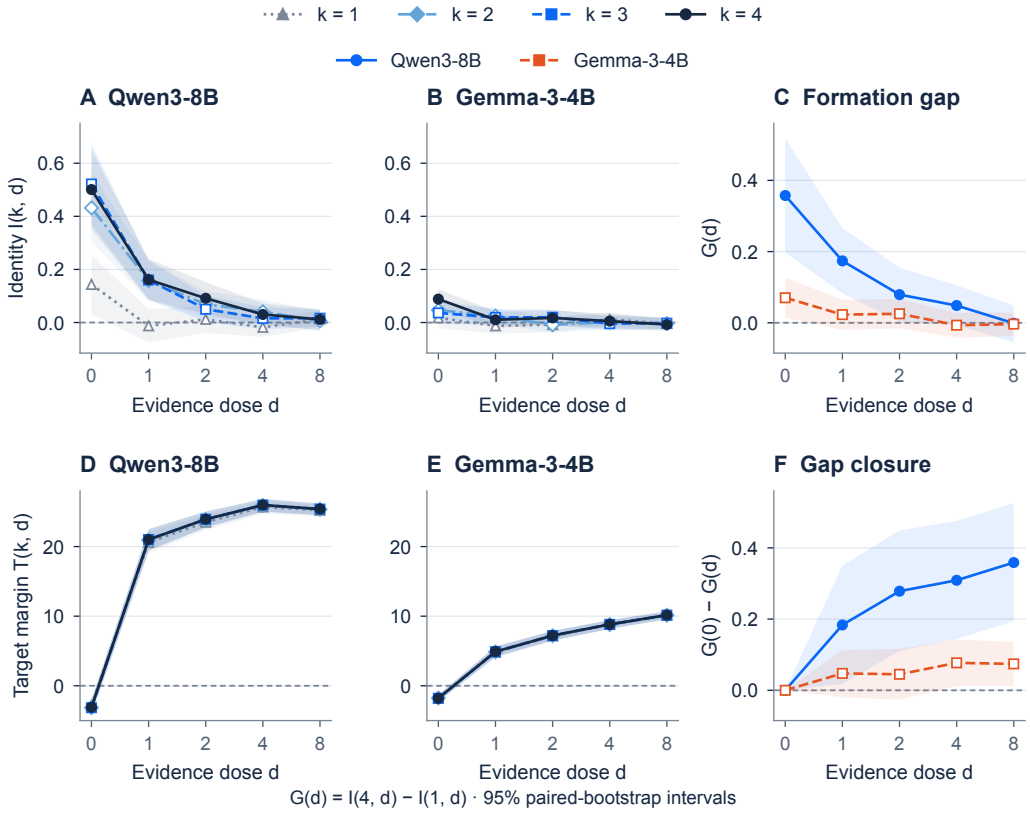

Figure 6: Complete coherent-usage-count by target-evidence surfaces for the temporal factorial. All 4×5 cells and intervals are shown with shared Qwen/Gemma axes within endpoint. The formation-gap panel defines $G(d) = I(4, d) - I(1, d)$ and marks zero; no local monotonicity claim is made.

reproduce the no-patch logits exactly in both models, confirming this structural no-op rather than a semantic absence.

Table 14: Worked Qwen3-8B example (frozen item 86, alpha_changed). English glosses abbreviate the released Chinese prompts. Source histories contain four changed-form and four unchanged-form usages; the unchanged concept is identical across self, donor, and random.

Source	Changed form s_a (one of four usage glosses)	Unchanged form s_b (held fixed)
Self	"After a week of overtime, the reports were almost unbearable."	"He lies on the sofa and has no interest in finding work."
Donor	"I suppressed a laugh as a colleague embarrassed himself."	same four usages
Random	"The new phone design is ugly and unergonomic."	same four usages

Table 15: Raw definition-readout log probabilities for the worked example at $c = 4, d = 0$. Here y_T, y_D, y_R, and y_F denote the answer labels for the changed form's original target concept, the donor concept, the matched random-control concept, and the unchanged form's fixed concept, respectively. These option roles remain the same for both queried forms. The margin is $m = \log p(y_D) - \log p(y_R)$, and $I = m(\widetilde{H}_D) - m(\widetilde{H}_R)$.

Query	Patched source	$\log p(y_T)$	$\log p(y_D)$	$\log p(y_R)$	$\log p(y_F)$	m
changed	donor	-4.265	-0.015	-7.015	-9.515	7.000
changed	random	-3.306	-0.056	-4.056	-7.306	4.000
unchanged	donor	-5.753	-0.003	-10.003	-11.753	10.000
unchanged	random	-4.759	-0.009	-7.759	-9.509	7.750

An independently defined uniform seven-useful-depth schedule is positive on all three endpoints in both model families. Qwen's identity, selectivity, and extension effects are respectively +1.681 ([+1.351, +2.033]), +0.523 ([+0.288, +0.756]), and +2.391 ([+2.067, +2.723]); Gemma's are +0.490 ([+0.319, +0.678]), +0.387 ([+0.236, +0.533]), and +0.567 ([+0.447, +0.693]). Simultaneous max-stat bands leave all three effects positive at Qwen layers 11 and 13–20 and Gemma layers 7–9. Thus the result is neither confined to the original depth list nor dependent only on repeated multi-depth injection.

No tested sparse budget through 12 met the discovery-frozen near-dense non-inferiority rule in either model, and equal-budget controls show that exact placement changes effect magnitude. The audit therefore establishes robust sparse causal access, including single-layer sufficiency for these three transport endpoints, but not a minimal or optimal layer set, a saturation frontier, necessary layers, or a storage location.

B.3 Boundary-span cross-model evidence arbitration

Table 19 reports the clean deployment panel at four source usages for 300 concepts per model. It evaluates donor identity, attenuation of donor influence, and increasing target margins under the boundary-span interface in Section 4.3, separately from the nonce-occurrence scale extension in Figure 7.

B.4 State-control construction and norm audit

For each target item i in the sorted same-fold registry, the donor is fixed at $i + 5$ and the random control at $i+10$ (cyclically). Both are real concept states, both are rebound to the target surface before capture, and target, donor, and random IDs are always distinct. Source and target tensors share seven relative depths, the tokenizer-specific 7/8-position boundary span, and full residual width. The exact unbatched audit reproduces every registered donor and random dose-zero margin. At every sampled depth, the donor-minus-random raw-state RMS and donor-minus-random displacement-from-self RMS CIs include zero.

To test activation energy directly, we construct, independently at each depth, $S_R^{\mathrm{nm}} = S_T + (S_R - S_T)\|S_D - S_T\|/\|S_R - S_T\|$. This leaves the random direction unchanged but matches the donor displacement norm over the entire span. Table 20 shows that donor selectivity survives, whereas norm matching alone has no detectable effect. Because S_R^{nm} is constructed, this test addresses displacement energy rather than every possible off-manifold property.

Table 16: NewTerm-English extension with human-reviewed materials. Reported intervals use 20,000 concept-level bootstrap resamples; reciprocal-pair clustering gives the same behavioral verdict. These rows test behavior and compact-state deployment, not the full two-form singleton protocol.

Endpoint	Qwen3-8B	Gemma-3-4B
q7 coherent−shuffle	+45.22 pp [+39.67 pp, +50.67 pp]	+35.33 pp [+30.22 pp, +40.44 pp]
Identity at $d = 0$	+0.52458 [+0.38457, +0.66333]	+0.52375 [+0.30458, +0.75208]
Identity slope	-0.15351 [-0.19468, -0.11270]	-0.15453 [-0.22102, -0.09040]
Evidence slope	+6.96942 [+6.68671, +7.24357]	+5.69496 [+5.30060, +6.08073]
Nonce−original identity	+0.25000 [+0.09457, +0.40625]	+0.29292 [+0.03167, +0.54917]
Pooled continuous anchor coefficient	+0.0361 [-0.0702, +0.1475]	

Table 17: **Held-out evaluation of the canonical single-layer interface.** One layer per model is selected on 50 discovery reciprocal-pair clusters using only identity, selectivity, and extension, then evaluated on 100 other target/donor pair clusters. Entries are effects with 95% paired pair-cluster-bootstrap intervals; all lower bounds exceed zero. The analysis has the locked-secondary status described in Section 2; all-role source sharing is detailed in Appendix C.3.

Model	Layer	Changed-form identity	Form selectivity	Label-free extension	Fixed-content formation	Evidence-driven revision
Qwen3-8B	15	**+1.088** [+0.854, +1.345]	**+0.572** [+0.398, +0.759]	**+1.455** [+1.239, +1.680]	**+0.313** [+0.163, +0.474]	**+0.207** [+0.054, +0.369]
Gemma-3-4B	11	**+1.087** [+0.884, +1.308]	**+0.161** [+0.056, +0.268]	**+1.573** [+1.392, +1.763]	**+0.163** [+0.046, +0.284]	**+0.124** [+0.011, +0.242]

B.5 Failed compact and complete-factorization accounts

A single pre-query boundary token fails all four older formation gates on Qwen3-8B, whereas its multi-position boundary span passes those gates. Under the stricter two-nonce construct test, however, even the complete compact boundary fails the joint lexical signature. Thus the old result establishes a compact causal summary for its task, not a sufficient lexical carrier. A separate full-complement factorial on Qwen3.5-2B did not provide evidence for a complete factorization; hence we do not identify an orthogonal complement with concept identity or claim a clean low-dimensional-interface / high-dimensional-identity decomposition.

B.6 Known-word confound sensitivity

Table 22 reports the principal imbalances and an adjusted sensitivity estimate. The regression uses observed paired gaps in baseline margin, surface token count, and target-prompt token count. Its intercept is the estimated nonce-minus-known identity gap when all observed gaps are zero; it is extrapolative when common support is sparse. Qwen3-8B has only 4/300 items within five margin units, Qwen3.5-4B has 238/300, and Gemma-3-4B has 217/300. The adjusted intervals cross zero for all three models. Threshold-matched analyses are mixed, so the raw Table 21 contrasts remain moderator evidence rather than a confound-free estimate.

B.7 Nonce-assignment robustness

The locked behavior includes three cyclic nonce/order schedules whose constituent-token multisets are identical. All pair-cluster binding-gap CIs are above zero and point estimates span only +41.50 pp–+42.50 pp. Separately, a no-usage diagnostic cycles the same three assignments for each concept and asks the model to choose among three target-aligned meanings. Pooled accuracy is 34.00% and its deviation from chance is +0.67 pp (Table 23). This finds no systematic target-aligned bias in the tested option sets, while leaving open weaker or non-target-aligned semantics of individual constituents.

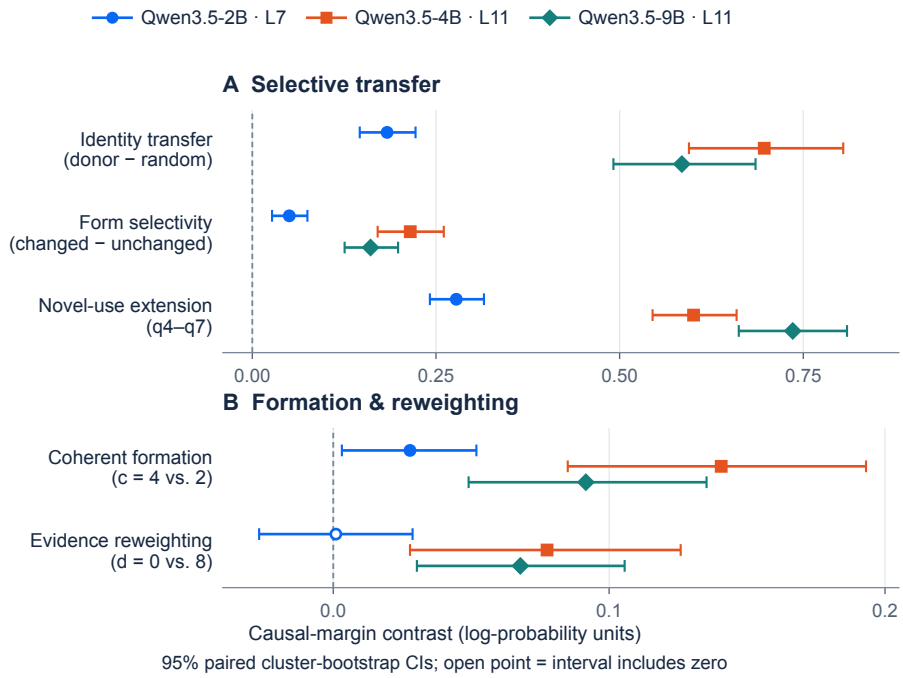

Figure 7: **Independent Qwen3.5 scale extension of the canonical singleton protocol.** One layer is selected per size from the same three discovery endpoints and tested on disjoint held-out clusters. All three sizes pass identity, selectivity, and donor-label-free extension; Qwen3.5-4B and -9B pass the full five-endpoint signature, whereas Qwen3.5-2B does not establish evidence-driven revision. Bars are 95% intervals from 20,000 paired base-pair-cluster bootstrap draws.

Table 18: Form-indexed usage-memory interchange (150 reciprocal pairs per model). Only the four blocks of the changed form are donor/random-swapped; unchanged-form blocks remain exact self states. Extension contains no donor definition or role label. Consistency c changes while source content, order, length, and form frequency remain fixed. All 95% paired cluster-bootstrap intervals pass the frozen directional gate.

Model	Changed-form identity	Form selectivity	Label-free extension	Fixed content $c{=}4 - c{=}2$	$c\times$ evidence interaction
Qwen3-8B	**+3.055** [+2.664, +3.461]	**+0.659** [+0.457, +0.867]	**+2.922** [+2.668, +3.181]	**+0.508** [+0.331, +0.693]	**+0.347** [+0.133, +0.559]
Gemma-3-4B	**+2.140** [+1.818, +2.483]	**+0.559** [+0.394, +0.729]	**+2.912** [+2.672, +3.157]	**+0.255** [+0.083, +0.431]	**+0.200** [+0.022, +0.382]

B.8 TANGENT CONTROLS

Rank-3 tangent spans use the symmetric splits q4+q5+q6→q7 and q4+q5+q7→q6. Each rank has same-fold cross-item, feature-shuffled, position-shuffled, and Gaussian norm-matched controls. Rank is fixed by the number of discovery queries. A rank-8 Partial Least Squares (PLS) basis recovers 94.7% of a local semantic-readout displacement but not selective output restoration; rank 32 yields the same verdict. A same-query output tangent has a positive local effect, but its cross-query CI crosses zero. These controls bound the tangent result to a local readout description and do not establish selective complete low-rank carriage under the tested interventions.

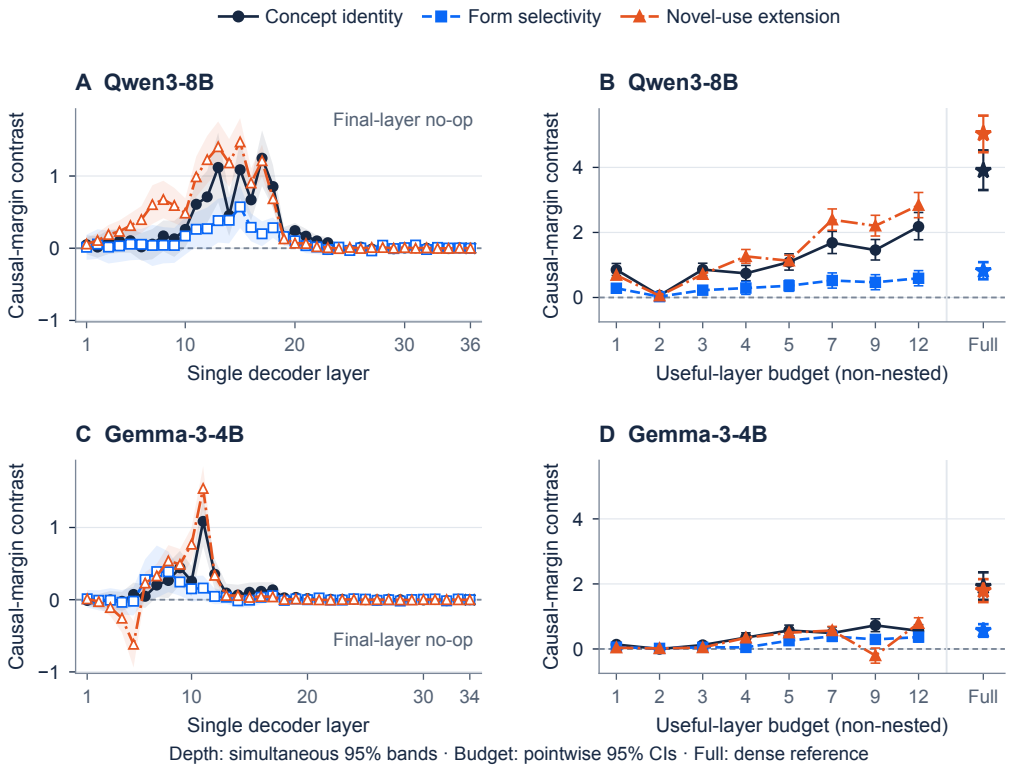

Figure 8: **Causal access is depth-structured but not tied to the registered layer list.** Single-layer profiles use max-stat simultaneous 95% bands; budget profiles use paired 95% intervals on the 100 held-out reciprocal-pair clusters per model. Budget schedules are independently defined, non-nested architecture-relative layer sets. "Full" replaces all useful layers as a dense reference, not a ground-truth reconstruction or an assumed upper bound. The exact final-layer no-op follows from causal order and serves as an implementation control.

Table 19: Boundary-span identity and evidence arbitration (300 concepts per model). Bold estimates have 95% CIs entirely in the prespecified direction. The lower block contains pre-specified capacity-boundary checkpoints.

Model	Portable identity $I(0)$: donor − random ↑	State influence β_I across dose ↓	Target evidence β_T across dose ↑	Joint verdict
Qwen3-8B	**+0.7896** [+0.5738, +1.0129]	**-0.2472** [-0.3118, -0.1834]	**+8.3384** [+7.9013, +8.7856]	**All three**
Qwen3.5-4B	**+0.0412** [+0.0233, +0.0596]	**-0.0128** [-0.0195, -0.0061]	**+1.5646** [+1.4840, +1.6448]	**All three**
Qwen3.5-9B	**+0.0562** [+0.0333, +0.0794]	**-0.0130** [-0.0205, -0.0055]	**+1.7414** [+1.6584, +1.8230]	**All three**
Gemma-3-4B	**+0.2733** [+0.1800, +0.3704]	**-0.0827** [-0.1101, -0.0552]	**+3.6571** [+3.4572, +3.8477]	**All three**
Qwen3.5-2B	+0.0087 [-0.0071, +0.0250]	+0.0002 [-0.0060, +0.0064]	**+1.2917** [+1.2155, +1.3671]	Evidence only
Phi-4-mini	+0.0079 [-0.0242, +0.0404]	+0.0002 [-0.0127, +0.0127]	**+2.6945** [+2.5274, +2.8606]	Evidence only

Table 20: Exact unbatched Qwen3-8B dose-zero state-control audit (300 concepts). The norm-matched control has the random direction and donor displacement norm.

Contrast	Estimate (95% CI)	Interpretation
Donor − real cross-concept random	**+0.7896** [+0.5742, +1.0079]	Registered control
Donor − norm-matched random	**+0.7854** [+0.5621, +1.0096]	Energy controlled
Norm-matched random − original random	+0.0042 [-0.0896, +0.1017]	CI crosses zero

Table 21: Matched nonce–known controls. Bold estimates pass the directional 95% CI gate. These are raw matched-concept contrasts, not performance- or tokenization-matched estimates.

Model	State reliance gap $I_N(0) - I_K(0) \uparrow$	Revision gap $\beta_{I,N} - \beta_{I,K} \downarrow$	Surface-anchor gap $\mathrm{Acc}_K(0) - \mathrm{Acc}_N(0) \uparrow$	Raw gates
Qwen3-8B	**+0.7206** [+0.5112, +0.9317]	**-0.2075** [-0.2700, -0.1451]	**+0.6233** [+0.5667, +0.6800]	**3/3**
Qwen3.5-4B	**+0.0281** [+0.0067, +0.0492]	**-0.0094** [-0.0176, -0.0012]	**+0.5500** [+0.4933, +0.6067]	**3/3**
Gemma-3-4B	**+0.2433** [+0.1387, +0.3510]	**-0.0763** [-0.1077, -0.0448]	**+0.2433** [+0.1733, +0.3133]	**3/3**

Table 22: Known-word confound sensitivity (300 paired concepts per model). The adjusted identity intercept controls observed baseline-margin, surface-token, and prompt-token gaps; every adjusted CI crosses zero.

Model	Exact surface-token match	Known − nonce baseline margin	Adjusted identity intercept
Qwen3-8B	29.00%	+28.613 [+27.273, +29.933]	+0.107 [-0.399, +0.618]
Qwen3.5-4B	29.00%	+3.641 [+3.444, +3.834]	-0.033 [-0.085, +0.018]
Gemma-3-4B	28.33%	+2.003 [+1.516, +2.486]	+3.591 [-1.831, +8.862]

Table 23: Qwen3-8B no-usage nonce-assignment diagnostic. Each concept is tested under three cyclic assignments; inference averages assignments within concept before 20,000 bootstrap resamples.

Endpoint	Estimate	95% CI
Target-aligned three-way accuracy	34.00%	[28.78%, 39.33%]
Accuracy minus $1/3$ chance	+0.67 pp	[-4.56 pp, +6.00 pp]

Table 24: Geometric transfer and causal restoration are separate endpoints.

Endpoint	Estimate	95% CI
Held-out tangent energy captured		
Within concept, rank 1	49.99%	–
Within concept, rank 2	60.94%	–
Within concept, rank 3	65.29%	[63.74%, 66.79%]
Cross item, rank 3	15.64%	–
Position shuffled, rank 3	3.85%	–
Local margin restoration under the tested contrast		
Within concept, rank 3	0.01666	[0.00261, 0.03101]
Rank-3 complement	0.02572	–
Within − cross item	0.00613	[-0.00902, 0.02116]

C History Formation, Common Attenuation, and Source Dependence

This follow-up adds one model experiment (E1) and two offline analyses of existing outputs (A1/A2). It retains the original endpoints and layer choices; the last-occurrence contrast is a separate diagnostic.

C.1 Fixed-last-usage history diagnostic (E1)

Source construction. Within each donor or matched-random source, both conditions retain the same eight sentences, their order, total token count, and four occurrences of each form. The $c = 4$ condition has four consistent assignments per concept. The $c = 2$ condition swaps form assignments in the earliest two pairs of usages and leaves the later two pairs unchanged. Thus the last occurrence of the changed form has identical local usage, nonce token IDs, absolute position, and most recent recurrence interval. All changed assignments precede the captured positions. A nonce activation has access to its own token and causal prefix, not to any sentence suffix or subsequent usage.

Three-vector intervention in a fixed background. Only the three full-width residual vectors at the changed form's last occurrence are supplied by the donor or matched-random source. The other 21 vectors in the 24-vector interface come from the same self-$c4$ cache in all four source-by-consistency cells. These comprise nine earlier-occurrence vectors of the changed form and twelve vectors of the other form. Target inputs are shared, and computation proceeds normally after patching. Random denotes the existing matched concept-source condition, not tensor noise. The self background itself carries semantic information, so the measured effect is conditional on that background rather than evidence that three vectors independently contain a complete lexical entry.

Readout and inference. The experiment retains Qwen3-8B layer 15 and Gemma-3-4B layer 11 under the original runner's indexing convention. It uses the existing 100 evaluation reciprocal pairs per model, both orientations, q4–q7 novel-usage queries, both queried forms, and $d = 0$ only. The design yields $100 \times 2 \times 2 \times 2 \times 2 \times 4 = 6{,}400$ scored conditions per model. Queries and orientations are averaged within each pair before 20,000 reciprocal-pair bootstrap resamples; intervals are 95% percentile intervals with the source-sharing scope in Appendix C.3. Equation 8 defines the primary contrast, and Table 2 reports both constituent history effects.

The changed-form history effect is positive in both models. Qwen's $T_{\text{history}} = +0.13906$ ([+0.07250, +0.20469]) supports an additional form-selective increment. Gemma's $T_{\text{history}} = +0.02484$ ([-0.02953, +0.07922]) does not establish this increment, although both of its constituent history effects are positive. The separate intervals do not constitute a direct test of the difference between model families.

Relationship to the original interface. The original canonical $c = 2$ condition rotates swap locations; E1 fixes swaps to the earliest two pairs to hold the last local usage and recent recurrence interval constant. The three-vector follow-up and the original twelve-vector changed-form intervention therefore differ in history schedule as well as intervention size; they are not a size-only capacity ablation. E1 evaluates history sensitivity with the last local usage held fixed, whereas the separate follow-up in Section 3.3 evaluates novel-usage selectivity using the canonical 24-vector single-layer intervention at $c = 4, d = 0$. The existing unchanged-form extension control in Table 12 uses the original multi-depth interface. No new model, layer, query subset, or history schedule was selected in response to the Gemma result.

C.2 Common attenuation of consistency surfaces (A1)

For each model and interface separately, the analysis reconstructs the $c \in \{2, 3, 4\} \times d \in \{0, 1, 2, 4, 8\}$ identity-margin surface from the original outputs. It fits

$$\widehat{I}(c, d) = a(c)g(d), \qquad \min_{a,g} \sum_{c,d} \left[I(c, d) - a(c)g(d)\right]^2, \quad \|g\|_2 = 1, \quad g(0) \geq 0. \tag{11}$$

All fifteen cells have equal weight. The fit is unconstrained apart from scale and sign normalization: it uses the original margin scale without logarithms, positivity constraints, or monotonicity constraints.

RMSE and residual energy describe the in-sample fit, not held-out prediction or a formal comparison of internal update algorithms.

Table 25: **A separable model closely describes the aggregate consistency-by-evidence surface.** Equal-weight rank-one least squares on 15 mean identity-margin cells. Residual energy is divided by the uncentered original surface sum of squares, not by behavioral variance. Single-layer and original multi-depth interfaces are fitted separately.

Model	Interface	Pairs	RMSE	Residual energy (%)	Observed Δ_{CE}	Fitted Δ_{CE}
Qwen3-8B	Single layer	100	0.01655	0.0633	+0.20750	+0.23789
Qwen3-8B	Multi-depth	150	0.02203	0.0191	+0.34708	+0.44871
Gemma-3-4B	Single layer	100	0.01411	0.0484	+0.12437	+0.11034
Gemma-3-4B	Multi-depth	150	0.01859	0.0207	+0.19958	+0.22646

Under a separable response, the interaction is

$$\widehat{\Delta}_{CE} = [a(4) - a(2)][g(0) - g(8)]. \tag{12}$$

It can be positive when stronger initial influence is attenuated by the same factor. Table 25 shows that this simple form closely describes the aggregate surfaces and reproduces the positive interaction sign in all four model/interface combinations. It describes the dominant mean response without assigning a unique computational implementation.

The residual-energy percentage is $100 \sum_{c,d}(I - \widehat{I})^2 / \sum_{c,d} I^2$; its denominator is not mean-centered variance. Fitting each pair separately gives mean residual-energy percentages of 13.9744% for canonical Qwen and 7.4960% for canonical Gemma. The very small aggregate residuals therefore do not imply that every item follows the same surface. Pair-level fits, heterogeneous mixture analyses, bootstrap intervals, and matched-100 subsets of the original multi-depth protocol are retained in the analysis export. A1 uses the original consistency schedule, not E1's fixed-earliest-swap schedule.

C.3 Source sharing and inferential scope (A2)

Reproduction and shared sources. An offline reconstruction of the two models under the canonical single-layer, original multi-depth, and compact protocols reproduces all six original primary summaries' point estimates and 95% intervals within 10^{-10}. Those frozen results are retained. This numerical check is distinct from whether the resampling units cover all sources of dependence.

Table 26: **Source sharing across all roles.** Units are reciprocal pairs for the canonical design and concept-indexed units for compact deployment. A concept is counted as reused if it occurs in more than one analysis unit in any role. Components connect units that share a concept.

Design	Units	Distinct concepts	Reused concepts	Components
Canonical, all pairs	150	300	247	8
Canonical, evaluation	100	255	126	21
Compact	300	300	300	5

Table 27: **All-role leave-one-concept-out sensitivity.** For each deletion, all units using that concept in any role are removed. The range is over recomputed point estimates, not a confidence interval. No single source-concept deletion reverses these canonical interaction estimates.

Model	Original Δ_{CE}	Range of Δ_{CE} after deletion
Qwen3-8B	+0.20750	[+0.15753, +0.23151]
Gemma-3-4B	+0.12437	[+0.10052, +0.15722]

The main target/donor reciprocal pairs do not overlap, but including random and fixed source roles reveals cross-unit concept reuse (Table 26). Across all roles, canonical discovery and evaluation share

114 concepts. The compact design has 5 source-sharing components, each containing 60 units. A pair-level split is therefore not a split that isolates every source concept.

What the intervals describe. The reported bootstrap intervals quantify resampling stability under the fixed materials and pairing schedule. They do not incorporate jointly resampling all source concepts, drawing new usages, or constructing new pairings. Reproducing the original intervals does not remove this distinction; source reuse alone also does not determine whether the intervals are too narrow. The source-sharing counts and sensitivity analysis make this scope explicit without replacing the frozen endpoints.

All-role leave-one-concept-out sensitivity. For each source concept, the analysis removes every unit that uses it in any role and recomputes the endpoint. Canonical Δ_{CE} remains positive under every such deletion on both models (Table 27). These ranges measure sensitivity to an individual concept, not corrected confidence intervals or a new significance test. The analysis export also contains corresponding E1, original multi-depth, and compact sensitivity outputs.

D Artifact Provenance

The original headline macro-backed statistics are generated from a checked registry mapping LaTeX commands to JSON pointers in released summaries. The build records a SHA-256 digest for every consumed result file; additional literal appendix tables are tied to their released analysis manifests. No numerical endpoint in the main manuscript is filled from a human assumption or copied from an optimistic audit file.

The history and reanalysis follow-up uses a separate result manifest and macro file, preserving the original 438 frozen macros. Its source report records SHA-256 hashes for the E1 and A1/A2 summaries and the analysis export. The writing bundle records the report-transcribed values, their source passages, and those summary hashes. Local checks validate transcription and consistency; the full follow-up analysis outputs remain in the separate export rather than being represented as raw data in the writing bundle.

Additional follow-up controls. The same-interface novel-usage follow-up in Section 3.3 retains the original 100 evaluation pairs per model, selected layers, source pairings, and changed-form q4–q7 prompts, and adds the corresponding unchanged-form queries. Changed- and unchanged-form scores are recomputed in the same execution environment, and the paired selectivity contrast is calculated entirely from these rerun outputs. The original canonical estimates are retained separately. The Gemma-3-4B cross-wiring follow-up in Section 3.1 uses 150 reciprocal pairs and the three original nonce schedules, comprising 3,600 scored conditions. Both follow-ups use 20,000 paired reciprocal-pair bootstrap resamples with 95% percentile intervals.