WHEN DO NEURAL STATES REALIZE THE SAME CAUSAL VARIABLE?

Anonymous authors
Paper under double-blind review

ABSTRACT

Mechanistic explanations often describe neural states in terms of high-level variables. For such descriptions to function as causal abstractions, the distinctions they omit must be irrelevant to the downstream counterfactual behavior of interest. Yet causal adequacy is usually assessed at a particular neural state, leaving open whether the same description remains adequate as computation unfolds. We study this question in a controlled retrieval task where direct and relational queries reach the same selected entity through different computations. The relevant semantic intermediate is fixed by construction, while we independently vary how the entity is selected, which value it is assigned, and where its complete assignment record appears. We find three stages with different causal organization: entity identity is sufficient at an early direct-query state under our interventions, later selection states also depend on assignment-record arrangement, and later states are organized by answer identity. State-restoration interventions connect these stages causally: an effect established at selection is later reversible through a downstream context-sensitive state. Meanwhile, states reached through the direct and relational procedures become increasingly interchangeable even as the distinctions required for causal sufficiency become finer. Thus causal granularity changes across computation even though the semantic intermediate is held fixed: the same high-level variable can be causally adequate at one stage and too coarse at another. Mechanistic explanations must establish not only which high-level variable a neural state corresponds to, but also over which part of the computation that variable remains causally adequate.

1 INTRODUCTION

Mechanistic explanations compress neural computation into task-level variables such as which entity has been selected or which answer is being formed (Geiger et al., 2025). Causal abstraction gives a principled criterion for such descriptions: an aligned high-level variable is faithful only when the distinctions it omits are irrelevant to the downstream counterfactual behavior under consideration (Geiger et al., 2021; 2025). Yet causal adequacy is usually assessed at a particular neural state, while computation unfolds through many. This leaves a temporal question: does a task-level description that is adequate at one point remain adequate as computation proceeds? Prior work shows that causal efficacy and representational invariance can dissociate across formats (Opielka et al., 2026), and that abstraction faithfulness can vary across inputs (Li et al., 2026). What is not yet established is whether the causal granularity adequate to a fixed semantic intermediate can change within a single unfolding computation. We test this directly: if two neural states correspond to the same known intermediate, does that variable remain sufficient to characterize their downstream causal role at every stage?

Activation patching tests the downstream effect of substituting one internal state for another (Zhang & Nanda, 2024). Successful substitution establishes functional substitutability under the tested intervention, but the downstream computation may remain sensitive to information omitted by the states' shared semantic label. We therefore separate three questions: whether two states have the same semantic referent, whether one procedure can reuse a state produced by another, and whether their shared high-level variable retains the distinctions needed to predict downstream causal behavior.

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.

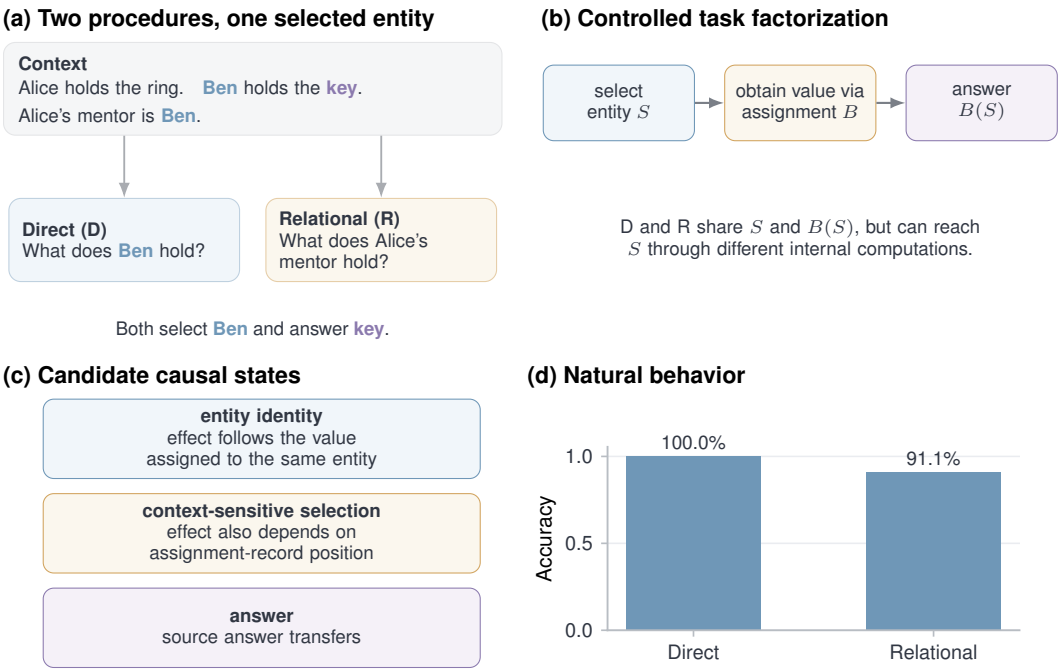

Figure 1: **Task and causal question.** (a) Direct and relational questions are paired to select the same entity through different computations. (b) Both procedures implement the same task factorization by selecting an entity, obtaining its value from the assignment map, and producing an answer. (c) The central question is whether the causal coarse-graining appropriate to the selected entity remains stable across computation: an entity-level abstraction may be adequate early, record arrangement may matter later, and a later state may become answer-like. (d) Both procedures are solved reliably enough for intervention analysis.

Testing causal sufficiency is difficult in ordinary language-model behavior because the semantic correspondence itself is usually inferred. We instead construct a retrieval setting in which the relevant intermediate is fixed by design. A context specifies entity–value assignments and one-hop relations between entities. A *direct* query (D) names an entity, whereas a *relational* query (R) reaches the same entity through a relation. Matched D/R questions share the selected entity and final answer while arriving at that entity through different computations. The task also lets us manipulate how an entity is selected, which value it is assigned, and where its complete assignment record appears while holding the semantic target fixed.

We test this with activation patching and counterfactual task variants that independently manipulate selection, assignment, and assignment-record arrangement. The central comparison keeps the source-selected entity and every entity–value fact fixed while rearranging complete assignment records. If that rearrangement leaves the effect of a patched state unchanged, entity identity is sufficient as the complete causal description under this intervention. If the effect changes even though the selected entity and its value remain fixed, entity identity alone is too coarse: downstream computation also depends on how that entity is situated in the assignment record. Separate assignment changes distinguish these selection states from later states whose causal effect is organized by the source answer.

Two findings organize the paper. First, the causally adequate description of the selected-entity intermediate changes across computation: entity identity is sufficient early in D, record arrangement matters later before retrieval, and answer identity becomes sufficient after retrieval. Restoration links these stages causally: an influence established at selection comes to depend on a downstream context-sensitive state. Second, D and R states become more reusable even as record sensitivity strengthens: they become more interchangeable while the distinctions required for causal sufficiency become finer. The central implication is temporal: a mechanistic abstraction must specify both the high-level variable and where in computation that description remains causally adequate.

2 BACKGROUND

Causal variables and interchange interventions. Causal abstraction formalizes when an interpretable high-level causal model is a faithful simplification of a lower-level system (Geiger et al., 2021; 2025). Interchange interventions test this relation by replacing aligned internal states and comparing the resulting counterfactual behavior; Distributed Alignment Search extends the same logic to learned subspaces (Geiger et al., 2024). For the sufficiency question studied here, a consequence of faithfulness is that distinctions omitted by the proposed high-level variable should not alter the relevant downstream counterfactual behavior. Recent work broadens or refines this framework in different ways: Xia & Bareinboim (2025) treats lossy representations, Pîslar et al. (2025) combine partial high-level models to capture input-dependent computation, and Li et al. (2026) partition the input space according to where a proposed interpretation is more or less faithful. Separately, Sutter et al. (2025) show that allowing unrestricted nonlinear alignment maps can make causal abstraction vacuous without additional representational constraints.

Shared structure across different computations. Prior work examines whether and at what level internal structure is shared across prompting formats, languages, modalities, and constructions (Opielka et al., 2026; Wu et al., 2025; Boguraev et al., 2025; Brinkmann et al., 2025; Nikankin et al., 2025; Tao et al., 2024). Different notions of sharing can dissociate: Opielka et al. (2026) separate causally effective function vectors from more format-invariant concept vectors, while Nikankin et al. (2025) find largely modality-specific circuits despite related functionality. A closely related approach, Hidden APIs, derives causal quotients from future-response signatures to identify reusable interfaces without researcher-specified latent labels (Ma et al., 2026). This literature motivates separating cross-context reuse from the information a faithful causal abstraction must preserve.

Entity binding and contextual retrieval. Entity retrieval provides a concrete setting in which contextual dependence can be studied directly. Language models must connect entities to attributes or values elsewhere in context, and prior work identifies structured binding mechanisms, effects of ordering interventions, relational organization, and mixtures of positional and lexical retrieval strategies (Feng & Steinhardt, 2024; Dai et al., 2024; 2026; Gur-Arieh et al., 2026). This work establishes that contextual organization can affect retrieval and provides controlled factors whose causal relevance can be varied while the retrieved content itself is preserved.

Positioning. These lines of work motivate two distinct questions: how the causal abstraction adequate to a known semantic intermediate changes as computation unfolds, and whether states produced by different procedures become interchangeable. Our setting separates them by fixing the selected entity independently of the neural states, varying contextual structure while preserving the underlying entity–value facts, and measuring cross-procedure reuse separately. This lets us test whether the same high-level variable can be causally adequate at one stage and too coarse at another even as D and R become increasingly interchangeable.

3 METHODS

3.1 TASK AND MODEL

Because each prompt comes from a structured task instance, the selected entity, its assigned value, and its record position are known independently of the neural activations. We can therefore vary these factors separately and test whether a task-level description is sufficient to predict a patched activation's downstream effect, and whether that description remains adequate as computation unfolds.

Task construction. Each world W contains four entities $\mathcal{E}$, four distinct values $\mathcal{V}$, a bijective assignment map $B : \mathcal{E} \to \mathcal{V}$, a one-hop relation map $f : \mathcal{E} \to \mathcal{E}$, and an arrangement $\rho : \{1, \ldots, 4\} \to \mathcal{E}$ of the four complete entity–value records, where $\rho(j)$ is the entity whose record appears in position j. A task instance $z = (W, p, Q)$ additionally specifies a query procedure $p \in \{\mathrm{D}, \mathrm{R}\}$ and query entity Q. The prompt presented to the model is a textual serialization $x = \sigma(z)$ of this structured instance.

For the two procedures, the selected entity and answer are

$$S_{\mathrm{D}} = Q_{\mathrm{D}}, \qquad S_{\mathrm{R}} = f(Q_{\mathrm{R}}), \qquad Y = B(S_p).$$

Thus D selects an entity explicitly, whereas R reaches one through a relation. Matched D/R instances satisfy $S_{\mathrm{D}} = S_{\mathrm{R}}$ and therefore have the same answer while reaching the selected entity through different computations (Figure 1). The shared selected entity is fixed by the task construction rather than inferred from the neural states.

A *task family* groups a sampled world with its prompt variants and controlled counterfactual instances. Statistical resampling keeps each family intact so related prompts from the same world are not treated as independent. Vocabulary sizes, prompt forms, generation details, and family construction appear in Appendix A.

Our detailed mechanistic analysis uses Qwen3-14B (Yang et al., 2025) with deterministic greedy generation and thinking disabled. Natural accuracy on the primary forms is near-perfect for D and 0.911 for R (Figure 1d); additional competence checks appear in Appendix A. The same broad causal sequence reproduces in independently localized Mistral-Small-24B and Phi-4 analyses (Appendix E). Intervention analyses include only prompts answered correctly by the unedited model, so the mechanistic results characterize the correctly solved subset.

3.2 Activation patching and counterfactual tests

Activation patching. Each intervention uses a *source* task instance z_s, which supplies the residual-stream activation being tested, and a *base* instance z_b, whose computation receives it. At a specified layer and semantic position—the *intervention site*—we replace the full base residual-stream activation with the corresponding source activation and continue the base computation (Zhang & Nanda, 2024). Source and base positions are located independently from their semantic spans because corresponding semantic roles can occur at different absolute token indices in D and R. The main analyses use three semantic positions: the *named-entity position*, the query token that explicitly names the selected entity in D; the *relation-endpoint position*, the final token of the relational query phrase in R; and the *final-prompt position*, the final prompt token before generation. These labels refer to semantic positions rather than fixed absolute token indices.

Causal regimes of interest. We focus on three causal regimes, corresponding to three ways an activation at a tested site can affect downstream computation. In an *entity-level* regime, the activation effectively specifies which entity downstream computation should use; the answer is obtained from that entity under the current assignment map. In a *context-sensitive selection* regime, the activation acts before the answer is fixed, but selected-entity identity alone is no longer enough to predict its downstream effect because the arrangement of the assignment records also matters. In an *answer-level* regime, the activation instead determines the answer strongly enough that the source answer follows the patch even when the selected entity is assigned a different value. We refer to the first two regimes as *pre-answer*: the particular output value has not yet been fixed by the inserted activation and remains determined downstream from the current assignment context.

Counterfactual tests. The three controlled tests distinguish these possibilities by changing how the source and base instances are paired (Figure 2). The patching operation is the same in every test: we insert the source activation at the corresponding site in the base computation and observe the resulting output. What changes is the source–base pair, chosen so that different causal interpretations predict different outputs.

Selected-entity change test. We first choose a source and base such that their selected entities differ while the assignment map and record arrangement are the same. After inserting the source activation into the base computation, we check whether the patched output is the value assigned to the source-selected entity rather than the base-selected entity. *Selected-entity transfer* is the fraction of patches that produce this source-selected-entity value. When it is high, we say that the activation exhibits *selected-entity control* at that site. This establishes a source-directed causal effect but does not distinguish pre-answer selection from direct source-answer control; the assignment test makes that distinction.

Assignment test. We separate these explanations by choosing a source and base with the same selected entity but different values assigned to that entity. If the inserted activation specifies which

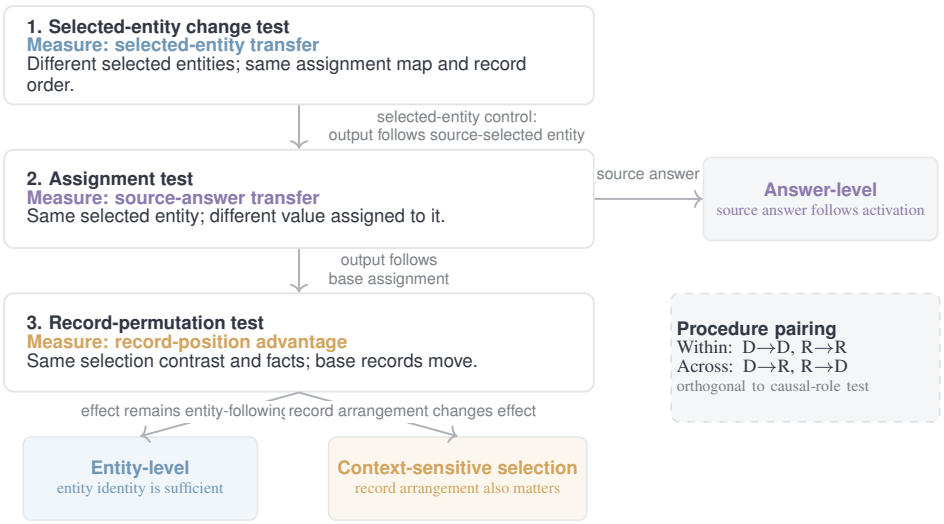

Figure 2: **Counterfactual tests for causal role at an intervention site.** The same activation-patching operation is applied to source–base pairs constructed to separate the three regimes. The assignment branch distinguishes answer-level from pre-answer effects; among pre-answer effects, record permutation distinguishes entity-level from context-sensitive selection. Procedure pairing is an orthogonal axis: each test can be applied within or across D and R.

entity downstream computation should use, downstream computation should use the base assignment map and produce that entity's new value. If the activation already determines the answer, the old source answer should instead remain despite the changed assignment. *Source-answer transfer* is the fraction of patches that preserve the source answer. Low source-answer transfer indicates that the activation has not yet fixed the answer and downstream computation determines it from the current assignment map. High transfer indicates an answer-level effect.

Record-permutation test. Low source-answer transfer leaves two pre-answer possibilities. The activation may specify the source-selected entity in a way that remains stable when records move, or its effect may also depend on the arrangement of the corresponding assignment record. Starting from a selected-entity-change source–base pair, we hold the source-selected entity, the source–base selection contrast, and every entity–value fact fixed while changing only the base record order. An *entity-following* effect continues to track the source-selected entity after the records move; a *record-position-following* effect instead tracks the entity occupying the corresponding source record position in the reordered base context. The *record-position advantage* compares these outcomes: negative values favor entity-following, whereas positive values favor record-position-following. More generally, attenuation of entity-following or increased corresponding-position following under permutation shows that record arrangement is causally relevant; positive record-position advantage is the stronger case in which the position prediction becomes dominant.

Within- and cross-procedure transfer. We write D→D and R→R for *within-procedure* transfers, and D→R and R→D for *cross-procedure* transfers. Cross-procedure transfer tests whether an activation produced by one procedure can perform the corresponding causal role in the other, rather than only sharing its semantic label. We evaluate the counterfactual tests under all four pairings at the relevant intervention sites; the semantic position can differ between D and R.

Complete counterfactual constructions and metric definitions appear in Appendix A and Appendix B, including the formal criterion for when two activations count as causally equivalent under this family of interventions.

3.3 Restoration and Recombination Interventions

The tests above characterize the causal role of an activation at each intervention site and whether activations can substitute across D and R. They do not show whether an earlier intervention exerts its effect through a later site, or whether a later activation can function when task information is supplied independently. We test these two properties with restoration and composition.

Causal dependence and restoration. To trace how an earlier activation's effect is carried forward, we use *restoration*. After an upstream source patch has propagated, we reset either the state at the source token or the downstream final-prompt state to its untouched base value at successive depths. Restoration attenuation measures the fraction of the upstream answer-changing effect removed by the reset. Comparing the two locations reveals whether the effect remains dependent on the source-position state or has become dependent on the downstream final-prompt state. Full definitions and the attenuation formula appear in Appendix D.1.

Systematic recombination and composition. A stronger test of a candidate intermediate is whether it can do its job when paired with task information that did not originally occur with it. We test this with *composition*: the assignment map and a later context-sensitive activation are manipulated independently, including combinations not produced together by the unmodified computation. *Composition success* measures whether the joint intervention produces the value predicted by the supplied activation under the independently supplied assignment map. Successful composition means that the activation can function with a new assignment context rather than only in the context in which it arose. Under record permutation, the same intervention tests whether its dependence on record arrangement remains active during recombination.

Restoration is applied to the selected D and R depth transitions. Composition is evaluated under all four procedure pairings on a separate evaluation population. Full assignment-map, restoration, composition, and analysis-population details appear in Appendix D.2, Appendix D.1, Appendix D.3, and Appendix B.

4 Results

We trace whether the causal description adequate for the selected-entity intermediate changes as computation proceeds. At each site, we apply the selected-entity-change, assignment, and record-permutation tests defined in Section 3.2 and summarized in Figure 2. In sequence, they establish selected-entity control, determine whether that effect is pre-answer or answer-level, and, among pre-answer effects, distinguish entity-level from context-sensitive selection.

4.1 Entity identity is initially sufficient, then becomes too coarse

D is the cleanest place to expect a simple causal description because the query supplies the selected entity directly; no one-hop relation is needed to determine which entity should be used. We begin with D→D patches at the named-entity position. Under the selected-entity-change test, patching the source activation redirects the base computation toward the source-selected entity through block 21 (Figure 3a), establishing selected-entity control in the operational sense of Section 3.2.

Block 20 exhibits the entity-level regime. The assignment test places block 20 on the pre-answer branch: the patched output follows the base assignment in 0.970 of cases and the source answer in 0.000. In the record-permutation test, the same source–base selection contrast and all entity–value facts are preserved while the base records move. At block 20, the effect remains with the entity (0.967 entity-following versus 0.001 record-position-following; Figure 3b). Thus block 20 is entity-level: the activation remains pre-answer, and entity identity is sufficient as the complete causal description for these interventions.

At block 21, the same description fails. Because D is given the relevant entity explicitly, and block 20 is well described by entity identity alone, the natural expectation is that this simple description might remain adequate until value retrieval. Block 21 violates that expectation. The assignment test keeps it on the pre-answer branch (0.923 base-assignment following versus 0.007 source-answer transfer), but the record-permutation test now yields 0.269 entity-following and 0.522 record-position-following (Figure 3b). Across the same families, record-position advantage changes by $+1.218$ $[1.181, 1.255]$ from block 20 to 21. This pattern places block 21 in the context-sensitive selection regime: the activation retains pre-answer selected-entity control, but entity identity alone no longer suffices as its complete causal description.

The simplest procedure does not remain causally simple. Even when the selected entity is supplied explicitly and the answer has not yet been fixed, its downstream causal role becomes dependent on

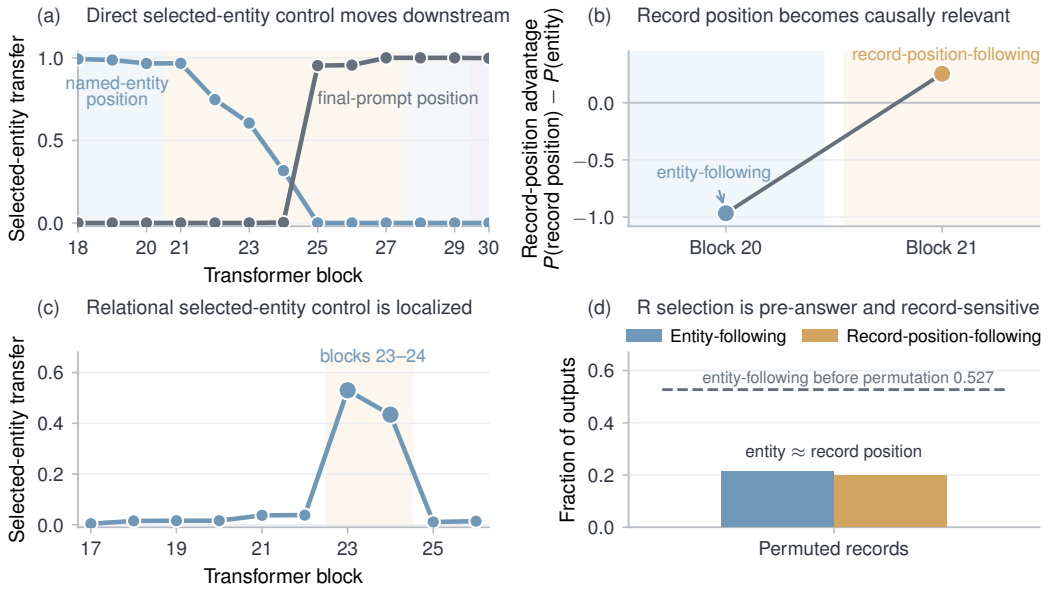

Figure 3: **The causal description of the selected entity changes across depth and procedure.** (a–b) D has selected-entity control at blocks 20–21; assignment keeps both pre-answer, while record permutation separates entity-level block 20 from context-sensitive block 21. (c–d) R first shows selected-entity control around blocks 23–24; at block 23 the effect remains pre-answer and becomes sensitive to record arrangement. Error bars are 95% task-family bootstrap intervals where shown.

record arrangement. Context sensitivity therefore cannot be attributed to relational lookup alone. The semantic intermediate remains the selected entity while the distinctions required for a complete causal description become finer across stages. The full-depth D profile shows a contiguous progression rather than isolated regime assignments (Appendix Figure 5).

4.2 The relational procedure reaches selected-entity control in a context-sensitive form

The D result leaves a natural alternative: perhaps selection generally passes through an entity-level phase before becoming context-sensitive. R provides the contrasting procedure because its selected entity is not named directly but must be determined through the mentor relation. Under the selected-entity-change test at the relation-endpoint position, R→R transfer first rises sharply around blocks 23–24 (Figure 3c), the first tested region where patching reliably redirects the computation toward the source-selected entity.

Relational selection remains pre-answer. At block 23, the assignment test shows that the patched output follows the base assignment in 0.396 of cases but preserves the source answer in only 0.093. The effect remains on the pre-answer branch rather than becoming answer-level.

Record arrangement already matters. The record-permutation test shows that entity identity alone is not sufficient at the same site. Entity-following is 0.527 before records are permuted and falls to 0.214 after permutation, while record-position-following reaches 0.198, giving record-order attenuation of 0.313 $[0.287, 0.340]$ (Figure 3d). Combined with the assignment result, this is the context-sensitive selection pattern.

At the tested sites, R does not reproduce D's earlier entity-level phase. The same selected entity becomes causally operative in different forms depending on how it is computed. Semantic identity therefore does not fix the complete causal description: its adequate granularity is stage- and procedure-dependent.

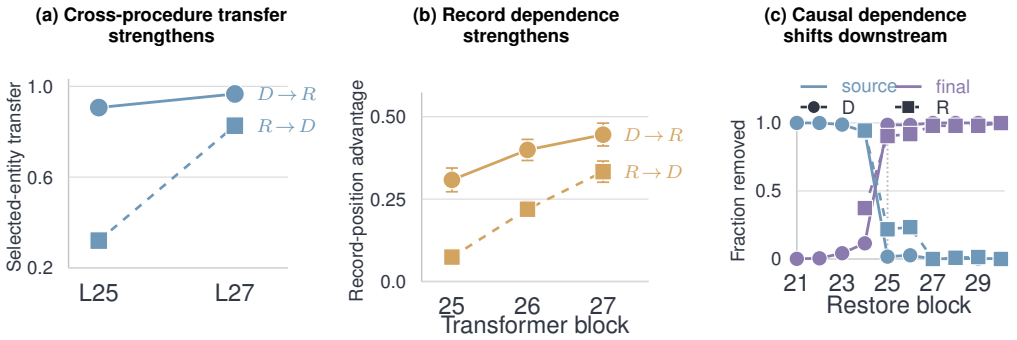

Figure 4: **Reuse strengthens while causal granularity becomes finer, and causal dependence shifts downstream.** (a) Cross-procedure transfer rises from block 25 to 27. (b) Record-position advantage rises over the same range. (c) Early in depth, restoring the source-position state removes the selection effect while restoring the final-prompt state does little; across blocks 24–25 this relationship reverses. Solid lines/circles denote D and dashed lines/squares denote R; color denotes the restored location. Error bars are 95% task-family bootstrap intervals where shown.

4.3 Cross-procedure reuse increases even as record-arrangement sensitivity strengthens

By the later final-prompt position, both procedures exhibit pre-answer, context-sensitive selection at the same semantic site. Sharing a site and causal regime, however, does not show that an activation produced by one procedure can substitute for one produced by the other. We test cross-procedure transfer using the D→R and R→D pairings defined in Section 3.2.

Cross-procedure transfer strengthens. Under the selected-entity-change test, transfer rises from 0.906 to 0.965 for D→R ($\Delta = +0.059$ $[0.048, 0.070]$) and from 0.320 to 0.827 for R→D ($\Delta = +0.507$ $[0.486, 0.529]$) between blocks 25 and 27 (Figure 4a). The two procedures become increasingly able to use one another's activations for the same source-directed causal effect.

If increasing transfer reflected convergence toward a coarser shared state, dependence on record arrangement should weaken as interchangeability improves.

Record dependence strengthens with it. The record-permutation test shows the opposite. Record-position advantage rises from 0.308 to 0.445 for D→R and from 0.074 to 0.333 for R→D over the same range (Figure 4b), with paired increases across adjacent depths. Cross-procedure reuse strengthens while the causal effect becomes *more*, not less, dependent on record arrangement. The states become more interchangeable while the distinctions required for causal sufficiency become finer.

4.4 The selection effect shifts from the source position to the final-prompt state

The preceding tests characterize individual depths but not how causal dependence shifts between them. We therefore restore either the source-position state or the final-prompt state after the earlier selection intervention has propagated. A crossover would show that the intervention's downstream effect has become dependent on a different neural state.

The crossover occurs in both procedures (Figure 4c). For D, source versus final restoration removes 0.935 $[0.924, 0.947]$ versus 0.115 $[0.101, 0.129]$ of the effect at block 24, but 0.017 $[0.012, 0.023]$ versus 0.986 $[0.981, 0.991]$ at block 25. R shifts from 0.944 versus 0.374 at block 24 to 0.218 versus 0.903 at block 25; by block 27 the corresponding values are approximately 0 and 0.979 $[0.967, 0.989]$. Thus the selection effect is initially reversible at the source state and later through a downstream state. The selected entity continues to constrain the answer, but the state whose restoration can reverse that effect changes as computation unfolds. Restoration therefore connects the depthwise regimes into stages of one evolving causal process rather than isolated classifications (Appendix D.1).

4.5 THE CONTEXT-SENSITIVE ACTIVATION RECOMBINES WITH AN INDEPENDENT ASSIGNMENT

By block 27, the selection effect depends on a downstream state whose causal role also depends on record arrangement. Could that state function only within the assignment context that produced it? We test this with the composition intervention from Section 3.3, using the pre-answer, context-sensitive block-27 final-prompt activation K across all four D/R procedure pairings.

Independent composition succeeds. We vary K and the assignment map independently (Appendix Figure 6a). Across D→D, R→R, D→R, and R→D, the joint intervention produces the predicted composed output at rates between 0.846 and 0.974 (Appendix Figure 6b). The activation can function with an assignment context different from the one that produced it.

Record dependence survives composition. Successful recombination could have reflected a simpler entity-level role under the new context. Instead, the record-permutation composition preserves the contextual effect: record-position-following is 0.643–0.788, compared with 0.119–0.153 entity-following (Appendix Figure 6c; Appendices D.2 and D.3). Thus the finer causal description at K does not prevent systematic recombination: the activation remains sensitive to record organization while functioning with an independently supplied assignment.

4.6 AFTER VALUE RETRIEVAL, ANSWER IDENTITY BECOMES SUFFICIENT

Composition places K before value retrieval: changing the assignment can change which value the supplied activation ultimately produces. Once computation has fixed the answer itself, the assignment test should reverse: changing the assignment should stop redirecting the patched output, and the source answer should instead follow the activation.

Returning to the assignment test from Section 3.2 at the block-30 final-prompt activation gives this expected pattern. Source-answer transfer is high across all four procedure pairings, ranging from 0.750 to 0.890 (Appendix C), so the activation's causal effect remains tied to the source answer when the assignment changes. This is the answer-level regime.

Block 30 marks the later transition from selecting an entity to fixing its value. The same broad sequence recurs in Mistral-Small-24B and Phi-4 (Appendix E). Across these sites, the central change is which distinctions a high-level description must preserve to predict downstream causal behavior.

5 DISCUSSION

Causal adequacy is stage-dependent, and causal dependence shifts downstream. Fixing the semantic intermediate does not fix the causal abstraction adequate to it. Entity identity is sufficient at an early D state, becomes too coarse before value retrieval, and later answer identity becomes sufficient. The restoration crossover connects these stagewise descriptions into one evolving computation: early, source restoration reverses the selection effect while final-prompt restoration does little; later, the relationship reverses. The selected entity remains the task-level intermediate, but the state through which it affects the answer changes. Thus the same semantic intermediate participates in successively different causal organizations: both the state through which it affects behavior and the distinctions needed to predict that effect change over time. A mechanistic abstraction is therefore a description of a phase of computation, not simply a label attached to a neural state; it should specify the variable, the distinctions that matter at that stage, and the part of computation over which that description remains causally adequate.

Granularity, reuse, invariance, and recombination are distinct. Opielka et al. (2026) show that causal efficacy across formats can coexist with format-specific representations, separating causal efficacy from representational invariance. Hidden APIs groups states by future computations and derives reusable interfaces from future-response signatures (Ma et al., 2026). Our results add a temporal causal-granularity axis: with the referent fixed, the distinctions required for causal sufficiency change even as cross-procedure substitutability increases, while the later context-sensitive state recombines with independent assignment information. Across models, the same semantic variable participates in different causal organizations as computation proceeds.

AI USE STATEMENT

In this work, we used generative AI tools to assist with writing, code generation, and generating synthetic datasets. All AI-assisted material was reviewed by the authors. AI-generated code and synthetic-data generation procedures were checked against the task specification and experimental tests before use. We take responsibility for the final content of this work, including text, code, and artifacts produced with the aid of generative AI.

REFERENCES

Allen Institute for AI. OLMo-2-0325-32B-Instruct. Model card, 2025. https://huggingface.co/allenai/OLMo-2-0325-32B-Instruct.

Sasha Boguraev, Christopher Potts, and Kyle Mahowald. Causal interventions reveal shared structure across English filler–gap constructions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 25021–25042, 2025.

Jannik Brinkmann, Chris Wendler, Christian Bartelt, and Aaron Mueller. Large language models share representations of latent grammatical concepts across typologically diverse languages. In *Proceedings of NAACL-HLT*, pp. 6131–6150, 2025.

Qin Dai, Benjamin Heinzerling, and Kentaro Inui. Cell-based representation of relational binding in language models. In *Proceedings of ACL*, pp. 47464–47524, 2026.

Qin Dai, Benjamin Heinzerling, and Kentaro Inui. Representational analysis of binding in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17468–17493, 2024.

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

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

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

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

Gemma Team. Gemma-2-27B-IT. Model card, 2024. https://huggingface.co/google/gemma-2-27b-it.

Yoav Gur-Arieh, Mor Geva, and Atticus Geiger. Mixing mechanisms: How language models retrieve bound entities in-context. In *International Conference on Learning Representations*, 2026.

Puyin Li, Jiyuan Tan, Ahmad Jabbar, Thomas Icard, and Atticus Geiger. Bucketing the good apples: A method for diagnosing and improving causal abstraction. arXiv:2605.02234, 2026.

SiYuan Ma, Yiqin Luo, Zhangji, Canran Xiao, Albert Gao, Wei Wang, Qiwei Wu, Xinran Li, Jinfeng Wei, and Qixin Zhang. Hidden APIs in language models: Discovering reusable causal interfaces from forked futures. arXiv:2607.27617, 2026.

Microsoft. Phi-4. Model card, 2024. https://huggingface.co/microsoft/phi-4.

Mistral AI. Mistral-Small-24B-Instruct-2501. Model card, 2025. https://huggingface.co/mistralai/Mistral-Small-24B-Instruct-2501.

Yaniv Nikankin, Dana Arad, Yossi Gandelsman, and Yonatan Belinkov. Same task, different circuits: Disentangling modality-specific mechanisms in VLMs. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Gustaw Opielka, Hannes Rosenbusch, and Claire Stevenson. Causality $\neq$ invariance: Function and concept vectors in LLMs. In *International Conference on Learning Representations*, 2026.

Theodora-Mara Pîslar, Sara Magliacane, and Atticus Geiger. Combining causal models for more accurate abstractions of neural networks. In *Proceedings of the Fourth Conference on Causal Learning and Reasoning*, volume 275 of *PMLR*, pp. 114–138, 2025.

Denis Sutter, Julian Minder, Thomas Hofmann, and Tiago Pimentel. The non-linear representation dilemma: Is causal abstraction enough for mechanistic interpretability? In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Junyi Tao, Xiaoyin Chen, and Nelson F. Liu. Inference and verbalization functions during in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 16394–16421, 2024.

Zhaofeng Wu, Xinyan Velocity Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim. The semantic hub hypothesis: Language models share semantic representations across languages and modalities. In *International Conference on Learning Representations*, 2025.

Kevin Muyuan Xia and Elias Bareinboim. Causal abstraction inference under lossy representations. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *PMLR*, pp. 68225–68235, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. arXiv:2505.09388, 2025.

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

A Task Construction and Counterfactual Definitions

A.1 Worlds, prompts, and analysis units

Each world contains four entities sampled without replacement from a vocabulary of 32 first names and four values sampled without replacement from a vocabulary of 32 concrete object nouns. The assignment map

$$B : \mathcal{E} \to \mathcal{V}$$

is a bijection between entities and values. A single relation type, rendered as *mentor*, defines a function $f : \mathcal{E} \to \mathcal{E}$ with no self-edges. We write $\rho : \{1, \ldots, 4\} \to \mathcal{E}$ for assignment-record order, where $\rho(j)$ is the entity whose complete entity–value record appears in position j.

For procedure $p \in \{\mathrm{D}, \mathrm{R}\}$, let Q_{D} and Q_{R} denote the entity named in the direct and relational questions, respectively. D selects the named entity directly, whereas R selects the mentor of the queried entity:

$$S_{\mathrm{D}} = Q_{\mathrm{D}}, \qquad S_{\mathrm{R}} = f(Q_{\mathrm{R}}), \qquad Y = B(S_p).$$

Matched D/R prompts satisfy $S_{\mathrm{D}} = S_{\mathrm{R}}$. All four assignment facts precede all four relation facts; assignment-record order and relation-sentence order vary independently. The four prompt forms are:

	Assignment record	Direct question	Relational question
1	"X holds the v."	"What does X hold?"	"What does X's mentor hold?"
2	"X carries the v."	"What does X carry?"	"What does X's mentor carry?"
3	"X has the v."	"Which item does X have?"	"Which item does X's mentor have?"
4	"X holds the v."	"Which item does X hold?"	"Which item does X's mentor hold?"

All relation facts use `X's mentor is Y.`. For each sampled world, the four query entities are crossed with the four prompt forms for both D and R; controlled source–base and counterfactual variants derived from that world remain in the same task family. On the 256-family Qwen competence population, direct accuracy is $4094/4096 = 0.9995$ and relational accuracy is $3733/4096 = 0.911$. A separate additional-wording competence check varies prompt wording while preserving the structured task semantics; relational accuracy is 0.923. A separate relation-change check changes the query-relevant mentor edge while retaining the task format, so the correct relational answer changes; accuracy is 0.916.

Prompts are passed to `Qwen/Qwen3-14B` through the tokenizer chat template as one user message, without a supplied system message, with `add_generation_prompt=True` and thinking disabled. The user message begins `Read the facts, then answer the question. Reply with only the item, as a single word.`; the serialized assignment and relation facts follow, then the D or R question. Decoding is unrestricted and greedy for at most eight new tokens.

A *task family* is the independent statistical unit: all query entities, wording variants, source–base comparisons, and counterfactual variants derived from one sampled world remain grouped during resampling. Intervention analyses require the relevant unedited source and base prompts to be answered correctly.

A.2 Counterfactual constructions

The three main tests use the same activation-patching operation but different source–base constructions.

Selected-entity change. Source and base have different selected entities while sharing the assignment map and record arrangement:

$$S_s \neq S_b, \qquad B_s = B_b, \qquad \rho_s = \rho_b.$$

The source-selected-entity prediction is

$$Y_{\text{entity}} = B_b(S_s),$$

whereas the untouched base answer is $Y_{\text{base}} = B_b(S_b)$. A positive patching effect establishes selected-entity control but does not distinguish pre-answer selection from direct transfer of the source answer.

Assignment test. Source and base select the same entity but assign it different values, with record arrangement fixed:

$$S_s = S_b = S, \qquad \rho_s = \rho_b, \qquad B_s(S) \neq B_b(S).$$

The two competing outputs are the value of S under the base map, $B_b(S)$, and the source answer, $B_s(S)$. Following the base map indicates a pre-answer effect; preserving the source answer indicates answer-level control.

Record permutation. This test starts from a selected-entity-change comparison and holds that source–base selection contrast and the entity–value facts fixed while changing only the base assignment-record order. Let

$$j_s = \rho_s^{-1}(S_s)$$

be the source record position occupied by the source-selected entity. Under a permuted base arrangement ρ_b', the entity occupying the corresponding record position is $\rho_b'(j_s)$. The competing predictions are therefore

$$Y_{\text{entity}} = B_b(S_s), \qquad Y_{\text{position}} = B_b\big(\rho_b'\big(\rho_s^{-1}(S_s)\big)\big)\,.$$

Across the aligned and permuted versions, the source-selected entity, the source/base selection contrast, and every entity–value fact are unchanged; only record arrangement changes. Y_{position} is an operational corresponding-position prediction and does not assume a literal ordinal representation inside the model.

For example, let Noah↦kite, Boris↦whistle, Henry↦lantern, and Vera↦brush. If the source selects Henry while the base selects Boris, then $Y_{\text{base}} =$ *whistle* and $Y_{\text{entity}} =$ *lantern*. If Henry occupies source record position 3 but Boris occupies base position 3 after permutation, the corresponding-position prediction is $Y_{\text{position}} =$ *whistle*, whereas the entity prediction remains *lantern*.

B Activation Patching and Measures

B.1 Patch operation and semantic positions

Let $h_{\ell,t}(x)$ denote the residual-stream output after block ℓ at token position t for prompt x. Let t_s and t_b be the source and base token positions corresponding to the same semantic intervention site. A full-state patch replaces the live base activation with the source activation at those independently located positions:

$$\widetilde{h}_{\ell,t_b}(x_b) = h_{\ell,t_s}(x_s),$$

after which computation continues on the base prompt. The main analyses intervene at the D named-entity position, the R relation-endpoint position, and the final-prompt position. The assignment-map intervention additionally uses the entity-name and value endpoints in the four assignment records. For multi-token spans, the last overlapping model token is used.

Within-procedure comparisons are D→D and R→R; cross-procedure comparisons are D→R and R→D. The semantic position is located independently in source and base prompts.

B.2 Causal equivalence and numerical measures

The main text uses causal-role labels as operational summaries of intervention patterns. The idealized relation underlying those labels is substitutability over a specified intervention family. Let $\mathcal{I}$ be

a set of interventions and $\mathcal{C}$ a set of base computations, and let $R_i(h, c)$ denote the downstream response when state h is inserted into computation c under intervention i. States h and h' are causally equivalent relative to $(\mathcal{I}, \mathcal{C})$ when

$$R_i(h, c) = R_i(h', c) \qquad \forall i \in \mathcal{I},\ c \in \mathcal{C}.$$

The experiments provide finite, directional evidence about substitutability over the tested family rather than proving equivalence outside it.

Let $\widetilde{Y}$ be the greedy output after patching.

Selected-entity transfer. For selected-entity-change pairs,

$$T_{p\to q}^{(\ell,t)} = P\Big(\widetilde{Y} = Y_{\text{entity}} \mid B_s = B_b,\ \rho_s = \rho_b,\ S_s \neq S_b\Big).$$

Here t indexes the semantic intervention site rather than one shared absolute token position; source and base token indices are located separately as above. High transfer means that the patch makes the base computation answer for the source-selected entity.

Assignment outcomes. For assignment-test pairs, source-answer transfer is

$$P\Big(\widetilde{Y} = B_s(S) \mid S_s = S_b = S,\ \rho_s = \rho_b,\ B_s(S) \neq B_b(S)\Big),$$

and base-map following is the corresponding probability of $\widetilde{Y} = B_b(S)$. High base-map following with low source-answer transfer supports a pre-answer interpretation; high source-answer transfer supports an answer-level interpretation.

Record-order effects. For the permuted member of a record-order pair,

$$\Delta_{P-E} = P(\widetilde{Y} = Y_{\text{position}} \mid \text{permuted}) - P(\widetilde{Y} = Y_{\text{entity}} \mid \text{permuted}).$$

Positive Δ_{P-E} means corresponding record position predicts the output more often than entity identity; negative values mean entity-following remains more common. Position dominance is therefore a stronger condition than contextual dependence.

To detect contextual dependence even when entity-following remains the more common outcome, we also compare aligned and permuted versions of the same selected-entity contrast:

$$A_{\text{order}} = P(\widetilde{Y} = Y_{\text{entity}} \mid \text{aligned}) - P(\widetilde{Y} = Y_{\text{entity}} \mid \text{permuted}).$$

Positive attenuation means that rearranging complete records weakens entity-following. Together with increased corresponding-position following, this shows that record arrangement changes the causal effect without requiring $\Delta_{P-E} > 0$. This distinction is important for the R block-23 state and for the cross-model comparisons.

The entity-level, context-sensitive selection, and answer-level regimes are defined in Section 3.2; the measures above are the quantities used to distinguish them.

B.3 Uncertainty and populations

Confidence intervals use 10,000 task-family bootstrap resamples with 95% percentile intervals, preserving all dependent observations from a family within each replicate.

C Qwen Supporting Evidence

The selected sites in the main text lie on coherent depth profiles rather than isolated peaks. Figure 5 shows the full Qwen progression, and the tables below give the selected-site causal quantities and positional sensitivity analyses.

Analysis	Families	Statistical structure
Qwen selected-site profile	256	Family bootstrap over eligible counterfactual comparisons.
Final-prompt blocks 25–27	224	Same source–base units evaluated at all three depths.
Assignment-map validation	192 and 256	Independent evaluation populations with the same intervention.
Composition	256	1498 eligible units receive all four assignment/state combinations.
Record-order composition	subset of same 256	641 units where entity and corresponding-position predictions differ.
Cross-model selected sites	256 per model	Separate model populations; no cross-model pooling.

Table 1: **Independent units and paired comparisons.** Family count is the number of independently sampled task worlds; repeated prompt variants from one family remain grouped during resampling.

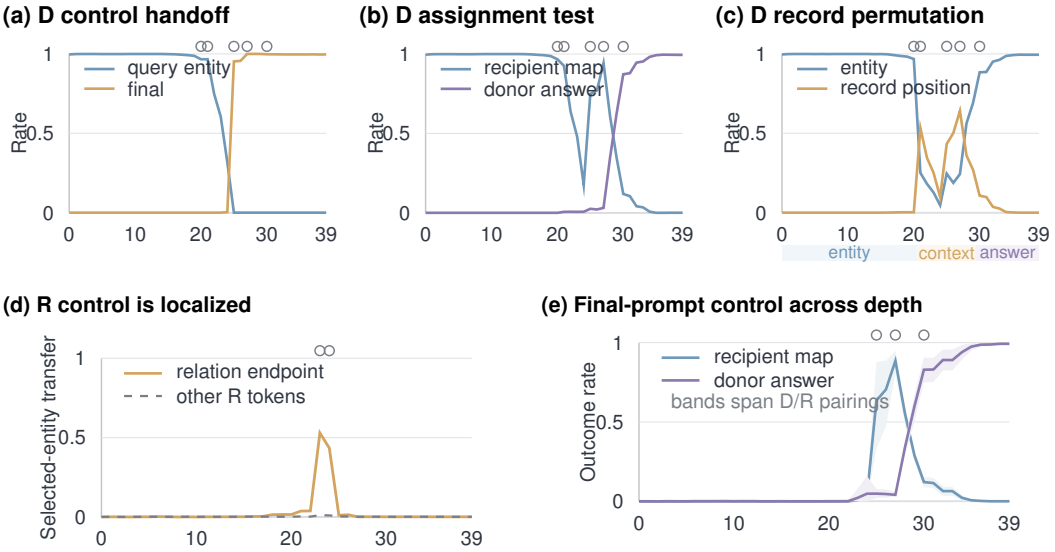

Figure 5: **Full-depth Qwen causal profile.** (a) In D, selected-entity control moves from the named-entity token to the final-prompt position. (b) The assignment test is recipient-map-like through the selection stages and becomes donor-answer-like after retrieval. (c) Under record permutation, entity-following dominates through block 20, record-position following becomes competitive or dominant through the context-sensitive region, and the later entity-following curve coincides with the donor answer under the fixed assignment. The strip gives the broad causal regimes. (d) R selected-entity control is sharply localized to the relation endpoint around blocks 23–24; the gray curve is the larger transfer rate at the other two R tokens. (e) At the final-prompt position, the four D/R pairings move from recipient-map behavior at blocks 25–27 toward donor-answer behavior from block 30 onward; lines are means and bands span the four pairings. Open circles mark the representative sites characterized in detail in the main text.

C.1 Full-depth causal profile

The D profile is contiguous. Entity identity accounts for the record-permutation effect through block 20; record organization becomes causally relevant from block 21 as named-entity control fades, and the final-prompt state carries the later context-sensitive regime before answer-level behavior dominates from block 30. R has a narrower selection window: selected-entity control is concentrated at the relation endpoint at blocks 23–24, where the semantic interventions identify context-sensitive selection. Across the four procedure pairings, final-prompt states follow the recipient map at blocks 25–27 and become donor-answer-like around block 30.

		Assignment test		Record-permutation test		
Procedure/site	Block	Base map	Source answer	Entity	Position	Attenuation
D named entity	20	.970	.000	.967	.001	.001
D named entity	21	.923	.007	.269	.522	.697
R relation endpoint	23	.396	.093	.214	.198	.313

Table 2: **Evidence separating the early causal regimes.** Base-map and source-answer columns are assignment-test outcomes. Entity and position columns are outcomes after record permutation. Record attenuation is aligned minus permuted entity-following. D block 20 is entity-level; D block 21 and R block 23 are context-sensitive selection. At R block 23, contextual dependence appears as strong attenuation even though entity- and position-following are comparable rather than position-dominant.

Pairing	Transfer L25	Transfer L27	Record advantage L25	Record advantage L27
D→R	.906	.965	.308	.445
R→D	.320	.827	.074	.333

Table 3: **Reuse and record dependence strengthen together.** Transfer values are selected-entity transfer. Record advantage is $P(\text{position}) - P(\text{entity})$ under record permutation. The paired transfer changes are $+0.059$ $[0.048, 0.070]$ and $+0.507$ $[0.486, 0.529]$ for D→R and R→D, respectively.

C.2 Selected-site causal regimes

Selected-entity-change interventions first establish selected-entity control; Table 2 gives the assignment and record-order quantities that distinguish entity-level, context-sensitive, and answer-level behavior at the early selected sites.

For D, the paired change in record-position advantage from block 20 to 21 is $+1.218$ $[1.181, 1.255]$. At R block 23, entity-following before permutation is 0.527, so the decline to 0.214 gives $A_{\text{order}} = 0.313$ $[0.287, 0.340]$. Reordering relation sentences instead of assignment records leaves entity-following at 0.569 while corresponding-position following is 0.094, tying the contextual effect to assignment-record organization rather than sentence order in general. The early D result concerns causal sufficiency at the named-entity intervention site; later states can retain entity information while requiring a finer complete description.

C.3 Cross-procedure reuse and contextual dependence

At the final-prompt position, selected-entity transfer increases between blocks 25 and 27 while record sensitivity strengthens on paired examples:

By block 30, source-answer transfer under assignment change is $0.874/0.785/0.890/0.750$ for D→D/R→R/D→R/R→D, respectively, giving the answer-level pattern reported in the main text.

C.4 Position-matched sensitivity

Equal-length variants fix the absolute intervention-token index in the within-procedure record-order test. Under this restriction, D block 20 remains entity-level (entity 0.993 $[0.987, 0.998]$ versus record position 0.001 $[0.000, 0.003]$), whereas block 21 remains context-sensitive (entity 0.312 $[0.282, 0.342]$ versus record position 0.517 $[0.485, 0.550]$).

At the shared later state, restricting donor and recipient final-prompt states to the same absolute token index preserves both high aligned transfer and record-position dependence at block 27 (Table 4). The within-procedure subsets retain nearly all families; exact cross-procedure matches are smaller because the D and R question forms differ in token length, so these estimates serve as a positional sensitivity analysis.

At block 25 the matched cross-procedure subsets retain the same directional record effect but with weaker separation, consistent with block 25 being an earlier, partially converged state. Block 27 provides the clearest shared context-sensitive interface.

Pairing	Aligned transfer all → matched	Entity	Position	Entity−position	Families A/M
D→D	.965 → .967 [.958,.976]	.204	.603	−.399 [−.444, −.351]	250/234
R→R	.864 → .862 [.845,.879]	.207	.557	−.349 [−.392, −.305]	250/241
D→R	.965 → .955 [.929,.978]	.200	.627	−.427 [−.563, −.279]	93/76
R→D	.827 → .847 [.783,.902]	.185	.476	−.292 [−.415, −.170]	58/44

Table 4: **Block-27 positional sensitivity.** "Matched" requires the donor and recipient final-prompt states to occur at the same absolute token index. Entity and position columns are the matched record-permutation outcomes; A/M gives aligned/mismatch family counts. Record position remains favored in every procedure pairing.

D Additional Causal Tests

D.1 Restoration across downstream depth

Restoration traces which neural state can still reverse an earlier selection effect as computation proceeds. After the upstream intervention propagates to block L, we reset either the residual state at the source token or the residual state at the final-prompt position to its untouched value. If p_{upstream} is the answer-changing rate after the upstream intervention and $p_{\text{restore}}(L)$ is the rate after restoration, we report

$$A(L) = 1 - \frac{p_{\text{restore}}(L)}{p_{\text{upstream}}}.$$

Larger $A(L)$ means that restoring that location removes more of the upstream behavioral effect.

Procedure	Restored state	L24	L25	L27
D, natural L20	source position	.935 [.924,.947]	.017 [.012,.023]	.001 [−.001,.003]
D, natural L20	final prompt	.115 [.101,.129]	.986 [.981,.991]	1.000 [.999,1.000]
R, relation L23	source position	.944 [.927,.960]	.218 [.187,.250]	.000 [−.009,.009]
R, relation L23	final prompt	.374 [.338,.411]	.903 [.879,.925]	.979 [.967,.989]

Table 5: **Restoration reveals a downstream shift in causal dependence.** Values are the fraction of the upstream answer-changing effect removed by restoring the indicated state. The source-position state dominates early, whereas the final-prompt state dominates from block 25 onward.

The two restoration locations exhibit complementary depth profiles. For D, source-position restoration removes nearly all of the effect through block 24 and almost none from block 25 onward, while final-prompt restoration shows the reverse pattern. R exhibits the same crossover over a broader transition: source restoration dominates at block 24, final-prompt restoration dominates at block 25, and the separation is nearly complete by block 27. Thus the state whose restoration reverses the earlier intervention changes across depth. The shift occurs where the final-prompt state enters the context-sensitive regime identified independently by the causal-state tests.

D.2 Independent assignment manipulation

Composition requires a causal handle on the assignment map that can be supplied independently of the later context-sensitive activation. We learn separate rank-32 entity-name and value subspaces at block 5 and replace only the source–base difference projected into the matching subspace at the eight assignment-record endpoints. For endpoint type $c \in \{E, V\}$, a learned matrix $W_c \in \mathbb{R}^{5120\times 32}$ is orthonormalized by a thin QR decomposition $W_c = U_c R_c$. The columns of U_c form an orthonormal basis for the learned rank-32 subspace, giving projector $P_c = U_c U_c^\top$. Let $c(t)$ denote whether token position t is an entity-name endpoint or a value endpoint. The edited base activation is

$$\widetilde{h}_t^{(b)} = h_t^{(b)} + P_{c(t)}\big(h_t^{(s)} - h_t^{(b)}\big).$$

Population	Changed-map targeting	Same-map preservation	Complete-map accuracy
192 families	.9988	1.000	.9953
256 families	.9987	1.000	.9949

Table 6: **Assignment-map intervention validation.** The same intervention is evaluated on two independent populations. A rank-matched random-subspace control produces no assignment-targeting successes on the corresponding eligible direct changed-map cases.

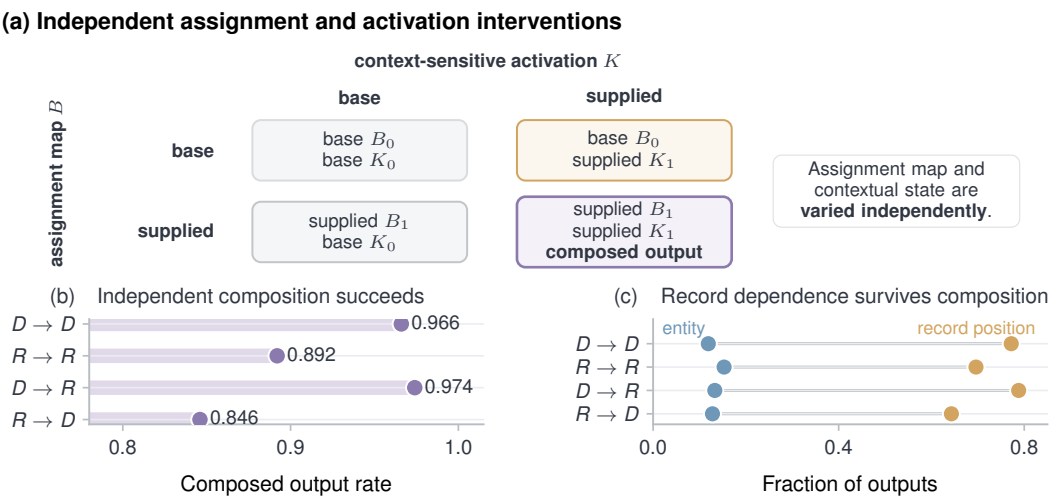

Figure 6: **Composition succeeds while context sensitivity persists.** (a) The assignment map B and context-sensitive activation K vary independently. (b) Joint interventions recover the predicted composed outputs across all four procedure pairings. (c) Under record permutation, record-position predictions continue to dominate entity predictions, showing that contextual dependence survives recombination.

The edit replaces only the component of the source–base activation difference lying in the learned subspace for that endpoint type; components orthogonal to that subspace remain as in the base activation. Only the two subspace matrices are optimized; model weights remain frozen. Training uses 384 task families with 75% changed-assignment examples and 25% same-assignment preservation examples, complete-answer teacher-forced negative log likelihood, Adam with learning rate 0.05, and batch size 32. Rank 32 is fixed before layer selection. A separate 192-family development split screens block outputs and selects block 5 and the retained checkpoint; the intervention is then frozen for an independent 192-family validation and the later 256-family confirmation. The rank-matched control uses independent Gaussian entity-name/value matrices with the same orthonormalization and projection-replacement operation. Changed-map targeting is the query-level rate at which the edit produces the value specified by the supplied assignment on eligible changed-map cases; same-map preservation is the corresponding query-level rate of retaining the correct output when source and base assignments already agree. Complete-map accuracy requires all entity queries associated with an eligible supplied map to agree with that map.

The high changed-map and complete-map rates show that the intervention controls the assignment map across entities rather than only steering one queried answer. Perfect same-map preservation and the random-subspace control support using it as an independent factor in the composition test.

D.3 Systematic recombination by composition

Composition tests whether the block-27 context-sensitive activation can function with assignment information supplied independently of the computation that produced it. Let B_0 be the base assignment and B_1 the supplied assignment; let K_0 be the base final-prompt activation and K_1 the supplied block-27 activation. Each eligible example receives all four combinations:

	K_0	K_1
B_0	(B_0, K_0)	(B_0, K_1)
B_1	(B_1, K_0)	(B_1, K_1)

The lower-right cell is the composition test. On aligned records, if K_1 supplies a usable selection-like intermediate, the predicted output is $B_1(S_1)$, where S_1 is the entity selected in the computation supplying K_1.

Across 1498 eligible examples, the joint intervention produces this composed output at rates 0.966 $[0.956, 0.975]$, 0.892 $[0.875, 0.908]$, 0.974 $[0.965, 0.982]$, and 0.846 $[0.826, 0.867]$ for D→D,

Pairing	Composition all → matched	Entity	Position	Entity−position	Families C/P
D→D	.966 → .963 [.951,.974]	.135	.730	$-.594$ [$-.677, -.511$]	255/174
R→R	.892 → .898 [.879,.917]	.195	.649	$-.453$ [$-.543, -.363$]	255/179
D→R	.974 → .972 [.942,.993]	.190	.738	$-.548$ [$-.765, -.314$]	89/37
R→D	.846 → .800 [.731,.865]	.180	.590	$-.410$ [$-.583, -.222$]	72/43

Table 7: **Position-matched composition sensitivity.** Matching requires the supplied and recipient block-27 final-prompt states to occur at the same absolute token index. Entity and position columns report the matched record-permutation composition panel; C/P gives composition/address-panel family counts.

R→R, D→R, and R→D. Recombination is therefore successful in every procedure pairing, with directional differences in efficacy.

Under record permutation, the source-selected entity and corresponding base record position predict different outputs. Let

$$X_1 = \rho_b\big(\rho_s^{-1}(S_1)\big)\,.$$

Thus X_1 is the entity occupying, in the base arrangement, the same record position occupied by S_1 in the source arrangement. The entity prediction is $B_1(S_1)$, while the corresponding-position prediction is $B_1(X_1)$. On 641 eligible examples, record-position-following is $0.772/0.696/0.788/0.643$ across D→D/R→R/D→R/R→D, compared with entity-following $0.119/0.153/0.133/0.128$. The resulting record-position advantages are $+0.65$ [$+0.60, +0.71$], $+0.54$ [$+0.49, +0.60$], $+0.66$ [$+0.60, +0.71$], and $+0.51$ [$+0.46, +0.57$].

A positional sensitivity analysis restricts the supplied and recipient K states to the same absolute final-prompt index. Composition remains high for D→D, R→R, and D→R, and remains substantial for R→D; record-position following continues to exceed entity-following in every pairing (Table 7).

Thus the supplied activation remains sensitive to record organization while functioning with an independently supplied assignment map. The composition result supports systematic reuse of a context-sensitive intermediate and remains stable under the positional sensitivity analysis. Its causal role is defined by the tested recombination and record-order effects rather than by a particular internal factorization.

E Cross-Model Recurrence of the Causal Sequence

Qwen exhibits a coarse progression from entity-level control to context-sensitive selection and then to answer-level control. That ordering could reflect Qwen-specific layer geometry rather than a recurrent organization of the computation. We therefore evaluate independently localized Mistral-Small-24B and Phi-4 models. The replication target is the ordering of causal regimes, not matching block numbers, transfer magnitudes, or transition spacing. All activation patches compare computations within the same model.

Causal behavior	Qwen	Mistral	Phi
D entity-level site	20	13	11
D contextual onset	21	14	12
D position-dominant crossover	21	15	12
R context-sensitive site	23	18	17
Shared final-prompt context-sensitive site	27	21	20
Answer-level site	30	30	26

Table 8: **Selected causal stages across models.** Entries are zero-indexed blocks. Mistral separates contextual onset from the later point at which record-position-following overtakes entity-following; Qwen and Phi show both at the same selected transition. The robust cross-model result is the coarse causal sequence, not invariant block depth or identical transition geometry.

E.1 Localization and task-applicability criteria

The models are localized independently so that replication does not depend on resemblance to Qwen depth. We evaluate Mistral-Small-24B-Instruct-2501 (Mistral AI, 2025), Phi-4 (Microsoft, 2024),

Gemma-2-27B-it (Gemma Team, 2024), and OLMo-2-32B-Instruct (Allen Institute for AI, 2025). Each checkpoint has an independent 256-family competence population. Models must exceed the prespecified minima below before causal localization. Passing models then use 192 new families for full-depth mapping and 256 further families for selected-site replication; all relevant unedited prompts must be correct before a counterfactual unit is retained.

Natural-task measure	Minimum	Mistral	Phi	Gemma	OLMo
Direct accuracy	.95	.9993	.9951	1.0000	.9871
Relational accuracy	.90	.9841	.9214	**.8579**	**.7332**
Joint D/R correctness	.80	.9834	.9182	.8579	**.7275**
Additional wording, D	.85	.9990	.9961	1.0000	.9863
Additional wording, R	.80	.9883	.9199	.8262	**.7363**
Relation change, D	.95	1.0000	1.0000	1.0000	.9990
Relation change, R	.90	.9893	.9238	**.8379**	**.7178**

Table 9: **Competence criteria for cross-model causal evaluation.** Bold entries fall below the required minimum. Mistral and Phi pass all criteria; Gemma and OLMo are retained only as task-applicability boundaries and are not assigned causal profiles.

Whole residual states are patched at every block output at the D named-entity position, R relation endpoint, or final-prompt position. The chosen sites and qualitative labels are fixed before the 256-family replication. For the paired assignment test, let F_B be the fraction of units that follow the base assignment in both assignment contexts and F_Y the fraction that preserve the fixed source answer in both. These paired events separate a pre-answer state that continues to use the current assignment map from an answer-level state that preserves the source answer.

A qualifying upstream selection site requires selected-entity control of at least 0.15, movement away from the natural base output of at least 0.10, a positive 95% interval for selected-entity minus source-answer following with at least 2:1 dominance, and pre-answer behavior under assignment change. The D entity reference is the last site before contextual onset that remains entity-dominant under record-order controls. Contextual onset is the first later qualifying site with positive 95% intervals for both entity attenuation and excess corresponding-position following,

$$X_{\text{order}} = P(\widetilde{Y} = Y_{\text{position}} \mid \text{permuted}) - P(\widetilde{Y} = Y_{\text{position}} \mid \text{aligned}).$$

Positive X_{order} means that permutation specifically increases corresponding-position following relative to the aligned baseline. Together with entity attenuation, it establishes record sensitivity without requiring the corresponding-position prediction to become dominant. Record-position dominance additionally requires a positive interval for record-position advantage. R uses the first contextual site with pre-answer assignment behavior.

At the final prompt, a contextual direction must also have at least 0.10 movement, positive selected-entity-minus-source-answer and $F_B - F_Y$ intervals, at least 2:1 selected-entity/source-answer dominance under aligned records, and a positive record-minus-entity interval with at least 2:1 record/entity dominance after permutation. The shared contextual site is the first block meeting those criteria in all four procedure pairings; the later answer-level site instead shows source-answer dominance under assignment change. These rules distinguish contextual onset from the stronger position-dominant crossover.

E.2 The early entity-to-context-sensitive transition recurs

Qwen D is entity-level before assignment-record arrangement becomes causally relevant. If that transition reflects a broader computational organization rather than a Qwen-specific block boundary, independently localized models should show an early entity-dominant state followed by a pre-answer state whose effect depends on record arrangement. Position-following itself does not have to become dominant at the same moment.

Mistral makes this separation explicit. In D, record-position advantage is -0.764 at block 13, -0.203 at block 14, and $+0.064$ at block 15. At block 14, entity-following remains more common than record-position-following, yet record rearrangement reduces entity-following by 0.453 $[0.433, 0.474]$ and increases corresponding-position following by 0.272 $[0.253, 0.290]$. Record arrangement is already causally relevant at block 14; the position prediction becomes dominant only at

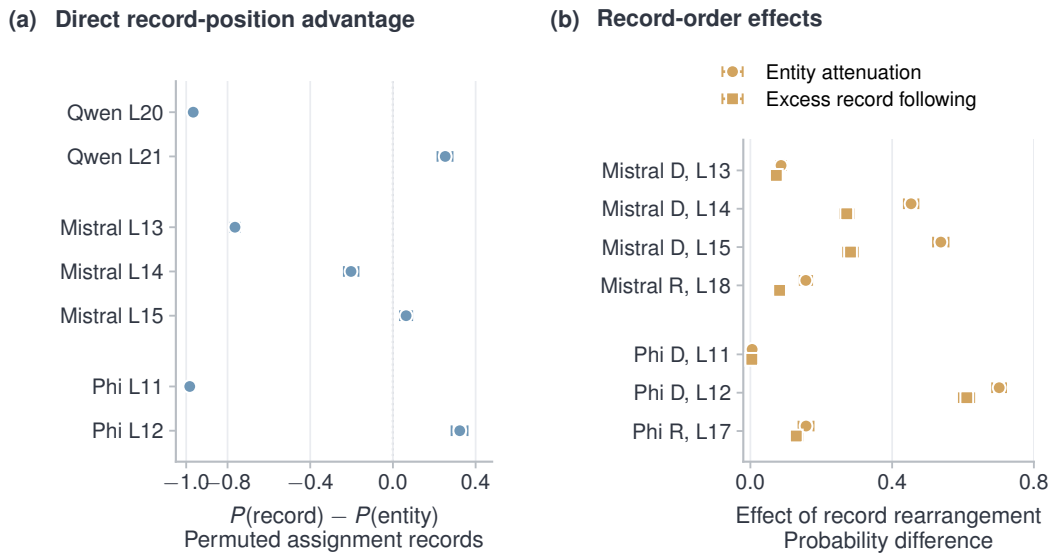

Figure 7: **Contextual dependence can precede position dominance.** Left: record-position advantage under record permutation. Right: aligned-to-permuted entity attenuation and excess corresponding-position following. Positive attenuation/excess establishes contextual dependence even when record-position advantage remains negative. Bars are 95% task-family bootstrap intervals at the selected sites.

block 15. Phi reaches contextual onset and position dominance together at block 12. This separation in Mistral confirms that the context-sensitive regime is defined by failure of entity identity alone, not by whichever of the entity or corresponding-position predictions has the larger rate.

The procedure contrast also recurs. At the selected R sites, assignment changes keep the effect pre-answer while record rearrangement attenuates entity-following by 0.156 $[0.140, 0.174]$ in Mistral and 0.157 $[0.136, 0.178]$ in Phi. D can exhibit an earlier entity-level state, whereas the first selected R state is already context-sensitive. The ordering matches Qwen even though the exact transition geometry differs across models.

E.3 The shared later context-sensitive regime recurs

Qwen later reaches a final-prompt state that is pre-answer and record-sensitive under all four D/R pairings. A recurrent shared regime should show the same causal classification in another model even if absolute transfer rates or directional asymmetries differ.

Mistral block 21 and Phi block 20 meet that criterion. Under aligned records, selected-entity transfer ranges from 0.664 to 0.760 in Mistral and 0.424 to 0.710 in Phi. Under record permutation, corresponding-position effects are present in all four pairings. The paired assignment test keeps these states on the pre-answer branch: in D→R, $F_B = 0.613$ in Mistral and 0.610 in Phi, compared with $F_Y = 0.087$ and 0.083.

The shared causal classification does not imply equal efficacy or symmetry. At the selected contextual sites, the descriptive difference $T_{D\to R} - T_{R\to D}$ is -0.027 $[-0.046, -0.008]$ in Mistral and $+0.225$ $[0.196, 0.252]$ in Phi. Even the direction of the asymmetry changes across models. What recurs is the tested causal organization of the state, not a fixed transfer magnitude or symmetry relation.

E.4 The sequence terminates in an answer-level state

If the full stagewise sequence recurs, a later final-prompt state should reverse the assignment-test pattern: changing the assignment should no longer redirect the patched output, and the source answer should remain tied to the supplied activation. This is the observed endpoint. In D→R, source-answer

transfer reaches 0.951 in Mistral at block 30 and 0.939 in Phi at block 26 (Figure 8c–d). These later sites provide the answer-level endpoint and complete the same coarse progression seen in Qwen.

E.5 Robustness and task applicability

The cross-model interpretation would be weaker if contextual onset were driven by token position, wording, or evaluation of models that do not reliably perform the task. Matching total token length and the patched token's absolute position preserves substantial contextual effects. At Mistral block 14, entity attenuation remains 0.424 and excess corresponding-position following 0.261; at Phi block 12 they remain 0.646 and 0.560. The exact Mistral block-15 position crossover is less stable under this restriction, whereas the broader contextual effect remains. Prompt wording likewise shifts some threshold-defined labels without removing the pooled record-arrangement dependence. These controls support the causal distinction more strongly than any exact block boundary.

The cross-model claim is restricted to models that solve the task reliably enough for the counterfactual comparisons. Mistral and Phi satisfy all competence requirements in Table 9. Gemma-2-27B and OLMo-2-32B do not, so they are treated as task-applicability boundaries rather than negative causal replications.

For Mistral and Phi, the selected causal sites are evaluated on independent 256-family populations with task-family bootstrap intervals. Across models that reliably solve the task, the replicated object is the ordering of causal regimes: an early D entity-level state, context-sensitive selection in D and R, a later context-sensitive state shared across procedure pairings, and a subsequent answer-level state. Absolute depth, transition spacing, transfer magnitude, and whether contextual onset coincides with position dominance remain model-dependent.

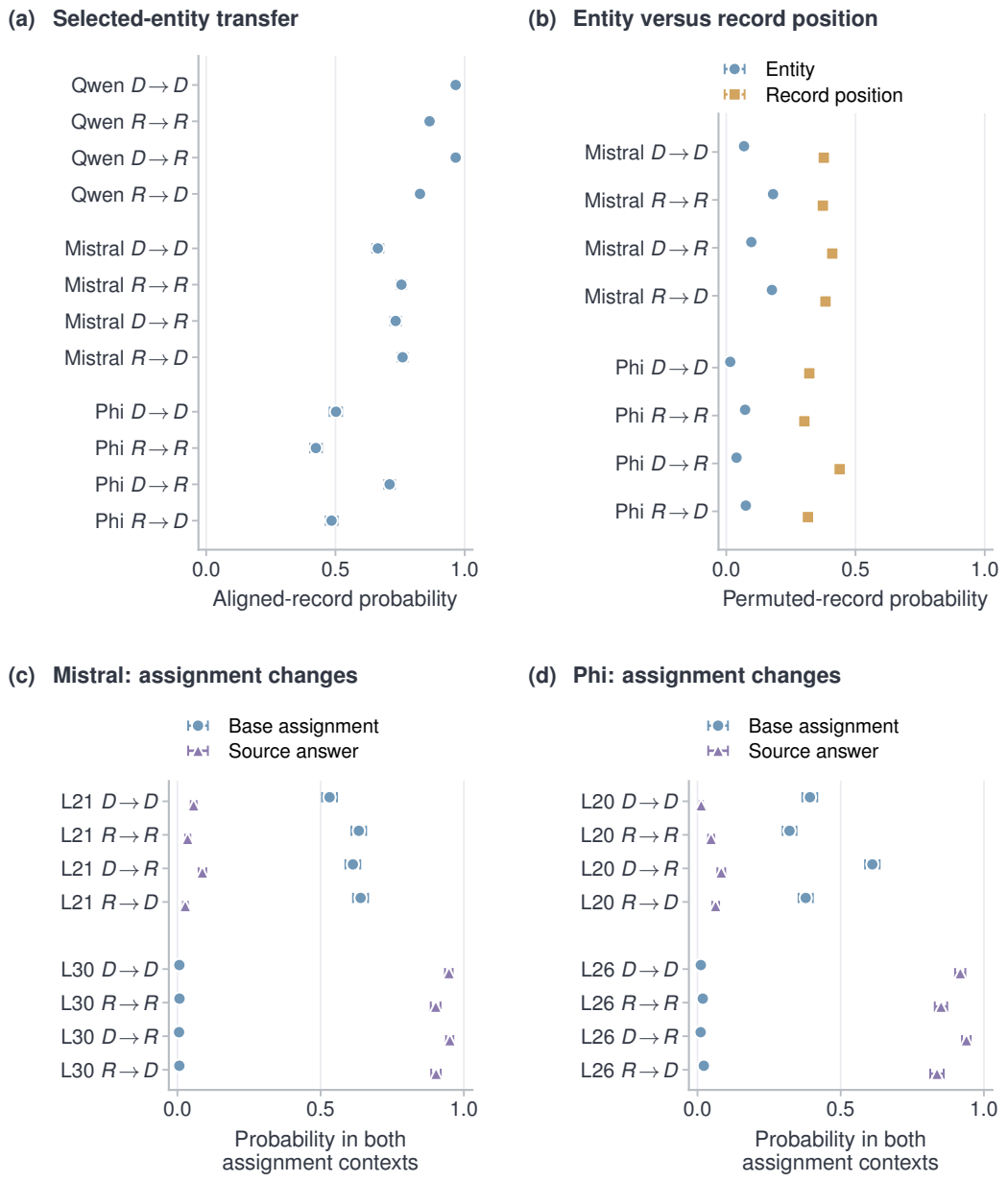

Figure 8: **The shared contextual regime recurs without fixed transfer strength.** (a) Selected-entity transfer at the later context-sensitive site. (b) Entity and corresponding-position following after record permutation. (c–d) Assignment changes distinguish the context-sensitive site from the later answer-level site: the earlier state follows the base assignment across contexts, whereas the later state predominantly preserves the source answer.