COMPOSITION AND CLOSURE: CAUSAL TESTS OF LINGUISTIC OPERATORS IN TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

A causal account of grammar in a Transformer must explain not only whether individual linguistic operations can be realized, but whether those operations continue to act correctly on states produced by earlier operations. We formalize this requirement as *complexity-bounded causal realization*: rank-bounded interchange maps are learned from unary counterfactuals at a fixed layer and span, then evaluated under held-out composition and on their *intervention closure*, the states reachable through valid intervention sequences. We show formally that perfect unary correctness on observational states does not imply correct composition, and give conditions under which closure-level one-step control bounds error along composed interventions. To study this distinction empirically, we introduce *OpClosure*, a synthetic benchmark with exact counterfactuals and controlled operator depth. Across independently trained Transformers, maps that perform well on unary interventions degrade systematically when applied to closure states, including after an earlier operation has been realized correctly. A modular reference is substantially more robust, while training with composed targets recovers much of the lost closure-level performance. We further observe analogous failures in controlled English experiments across pretrained models from multiple families and scales. These results suggest that unary causal success is insufficient evidence for compositional grammatical mechanisms: causal realization should be evaluated on the states that the proposed mechanism itself produces.

1 INTRODUCTION

Changing number in *The pilot is waiting* gives *The pilots are waiting*, changing tense gives *The pilot was waiting*, and applying both gives *The pilots were waiting*. Grammar specifies these transformations, their validity conditions, and composition (Montague, 1970; Lambek, 1958; Goldberg, 1995). Passivization requires a suitable clause frame. A causal account must explain whether each neural implementation still performs its operation after another has changed its input state.

Causal abstraction and distributed alignment search connect neural-state changes to high-level interventions (Geiger et al., 2021; 2024), including linguistic features (Arora et al., 2024). Interventional tests distinguish causal use from decodability (Elazar et al., 2021), while successful subspace patches can activate dormant pathways (Makelov et al., 2024). Because learned alignments can supply computation, bounding their class makes the explanatory budget explicit (Sutter et al., 2025).

A finite counterexample shows why reuse needs testing. A variable and its copy agree throughout the observational support, making both unary implementations exact. The first intervention changes the variable but leaves the copy stale. The second reads that copy and fails (Proposition 1, Figure 2). We therefore evaluate the *intervention closure*, states reached by valid sequences of the fixed interventions.

Existing work supports serial interventions, subspace composition, and typed-variable training separately. Our advance is their joint evaluation for linguistic operators. Building on recursive interchange (Geiger et al., 2025), *complexity-bounded causal realization* tests rank-bounded maps fitted independently on unary counterfactuals, frozen at a common layer and span, reused for held-out sequential composition, and evaluated on intervention-reached states (Definition 1).

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.

OpClosure provides exact counterfactuals, matched sources, and effective-depth control, so evaluated endpoints require the stated number of operations. Frozen interventions in standard synthetic models lose accuracy under composition and on reached states relative to unary tests. Descriptive calibration against the unary-trained modular reference identifies the model-attributable excess under ordinary training. Correct composed targets in model training raise fitted-map closure-local accuracy relative to that reference on the same replication dataset, closing most of the gap. Controlled English tests on seven pretrained models spanning three families and 360M to 3B parameters show that rank-bounded interchanges realize number and tense and their compositions. Per-operator scores distinguish preserved local accuracy from offsetting gains and losses (Figure 1).

We make three contributions. First, the realization criterion jointly tests linguistic operators with explicit intervention budgets and reuse domains (§3). Second, a machine-checked finite counterexample separates unary and composed correctness. Uniform closure-level control bounds expected composition error under Lipschitz operators and validity at decoded and ideal states (Theorem 1). We establish tightness and give per-assumption counterexamples for this stochastic bound (§4). Third, we provide OpClosure, measure and calibrate reuse losses, reduce them through composed model training, and establish rank-bounded realization in pretrained models (§5–§6).

2 Related work

Causal interventions on linguistic representations. Causal abstraction and distributed alignment search (DAS) relate neural states to high-level variables through interventions (Geiger et al., 2021; 2024), while recursive interchange permits interventions on states produced by earlier interventions (Geiger et al., 2025). In language models, CausalGym and RAVEL evaluate causal control and feature isolation (Arora et al., 2024; Huang et al., 2024), and prior mechanistic studies analyze phenomena such as agreement, entity binding, and filler–gap dependencies individually (Finlayson et al., 2021; Feng & Steinhardt, 2024; Boguraev et al., 2025). These works establish causal tests for individual representations or phenomena. We ask whether the *same* learned intervention remains valid after another intervention has changed the state on which it acts.

Composition and intervention-based explanations. Compositional structure has been studied through circuit composition, function-vector arithmetic, subnetworks, and reusable causal interfaces (Mondorf et al., 2025; Todd et al., 2024; Lepori et al., 2023; Ma et al., 2026); sequential interventions can also interfere when combined (Kolbeinsson et al., 2025). At the same time, intervention-based explanations require care: expressive alignments can supply computation themselves (Sutter et al., 2025), activation patches can recruit dormant pathways (Makelov et al., 2024), and conclusions depend on the patching procedure and metric (Zhang & Nanda, 2024; Méloux et al., 2026). We therefore bound the intervention class, fit operators independently on unary counterfactuals, and freeze them before compositional evaluation.

Reached states and compositional guarantees. Composition introduces a further requirement: later operators act on states created by earlier ones. Work on divergent representations studies the geometry of intervention-produced states (Grant et al., 2025), while approximate causal abstraction analyzes how local error propagates through composition (Beckers et al., 2020; Rischel & Weichwald, 2021; Lorenz & Tull, 2026). Formal approaches similarly seek guarantees that mechanistic explanations preserve intended behavior (Palumbo et al., 2025; Hadad et al., 2026). We make this reached-state requirement explicit through the *intervention closure*, evaluating frozen rank-bounded linguistic maps both under held-out sequential composition and on the states they themselves generate. Theorem 1 places its one-step premise on this closure. OpClosure instantiates the test with typed linguistic operations and exact counterfactuals; broader connections to formal linguistics and Transformer expressivity appear in Appendix K.1.

3 A criterion for causal realization of linguistic operators

3.1 Linguistic operators and bounded neural implementations

In OpClosure, NUM, TENSE, and POL toggle number, tense, and polarity. Passivization (PASS) requires an active transitive frame without an open gap. States encode role bindings, clause features, construction type, and filler-gap status in finite state spaces $\mathcal{Z}_\tau$ indexed by a finite set $\mathcal{T}$ of types,

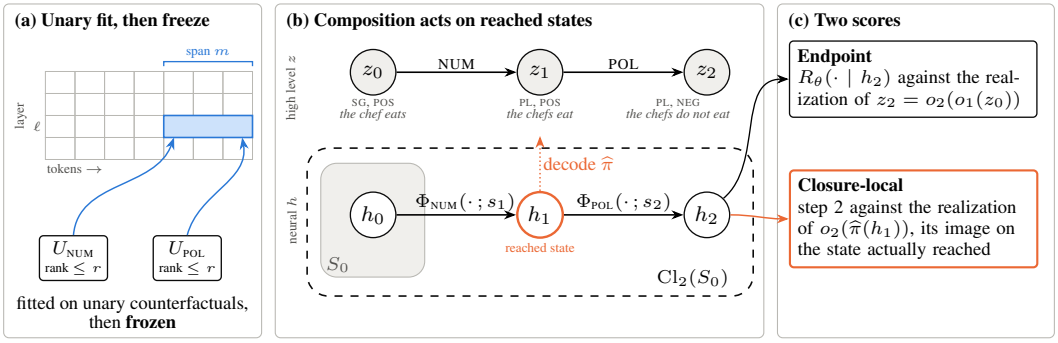


Figure 1: **Evaluating the reuse of linguistic interventions.** Rank-bounded number and polarity maps are fitted on unary counterfactuals and frozen at a common layer and span. Sequential application produces h_0, h_1, h_2, with the second source s_2 matched under the declared running-state policy. The sentences illustrate high-level feature changes. The right-hand comparison scores the final neural outcome against the realization of $z_2 = o_2(o_1(z_0))$. Closure-local scoring uses the realization of $o_2(\widehat{\pi}(h_1))$ as its target, where $\widehat{\pi}(h_1)$ is the decoded reached state.

with $\mathcal{Z} = \bigsqcup_{\tau \in \mathcal{T}} \mathcal{Z}_\tau$. Each operator in $\mathcal{O}$ is a partial map $o : \mathcal{Z}_\tau \rightharpoonup \mathcal{Z}_{\tau'}$ with validity domain $\text{dom}(o) \subseteq \mathcal{Z}_\tau$. The inventory includes controlled THEME-role reassignment (BIND), WH-extraction, and caused-motion introduction (Goldberg, 1995) (Appendix C.3). A *word* $w = o_k \circ \cdots \circ o_1$ applies $o_1, \ldots, o_k$ sequentially and is valid when each operator is defined at the state it receives. $\mathcal{W}_{\le K}$ contains valid words of length at most K. Validity depends on order. Active benchmark operators commute wherever both orders are valid (Appendix K.2).

For a trained Transformer T_θ with frozen weights, let $h_t^\ell = H_\theta^\ell(x_{\le t}) \in \mathbb{R}^d$ be the residual activation after block ℓ at token t of prefix $x_{\le t}$. Candidate implementations for all active operators share a validation-selected layer and right-aligned span with nominal maximum look-back budget m. At site t^*, the block width is $q_x = \min\{m, t^* + 1\}$. Base and source positions align by offset from their respective endpoints (boundary and source-edge rules in Appendix A.1). At each offset, h_b is the base activation and h_s the source, or donor, activation. The donor matches unchanged high-level fields and supplies the value prescribed by o (Appendix D.1). At every aligned position, we apply the interchange map with the same low-rank projector

$$\Phi_o(h_b, h_s) \;=\; h_b + P_o\,(h_s - h_b), \qquad P_o = U_o U_o^\top, \quad U_o^\top U_o = I_{r_o}, \tag{1}$$

where $U_o \in \mathbb{R}^{d \times r_o}$ has orthonormal columns and $r_o \le r_{\max}$. This preserves $(I - P_o)h_b$ from the base and replaces its projected component with $P_o h_s$ from the donor. Rank bounds writable directions per position, while m bounds affected positions, controlling alignment flexibility (Sutter et al., 2025). Appendix D.2 gives the parameterization and fitting details.

Each U_o is fitted independently to matched unary counterfactuals in the training pool. Validation selects the common layer and nominal span, per-operator ranks, and decision thresholds. Maps, settings, and decision rules are frozen before evaluation on held-out lexical items, operator pairs and triples, and test examples. For a word $w = o_k \circ \cdots \circ o_1$, we reuse the maps sequentially

$$\Phi_w(h_b;\, h_{s_1}, \ldots, h_{s_k}) \;=\; \Phi_{o_k}\big(\cdots \Phi_{o_2}\big(\Phi_{o_1}(h_b, h_{s_1}),\, h_{s_2}\big) \cdots,\, h_{s_k}\big). \tag{2}$$

Each map receives the preceding intervened activation directly. Donors match the running high-level state under the declared policy (Appendix D.1), so reuse is tested on activations earlier maps created.

3.2 Reached States and Observable Outcomes

Sequential reuse generates a domain of reached activations. Let $S_0 \subseteq \bigsqcup_{q=1}^m (\mathbb{R}^d)^q$ be the boundary-aware blocks realized by the un-intervened evaluation distribution over prefixes, the *observational support*. Valid words of length at most K with admissible observational sources generate the *K-step intervention closure*

$$\text{Cl}_K(S_0) \;=\; \big\{\, \Phi_w(h;\, h_{s_1}, \ldots, h_{s_k}) \;:\; h \in S_0,\; w \in \mathcal{W}_{\le K} \text{ valid for } h,\; h_{s_i} \in S_0 \,\big\}, \tag{3}$$

including starting states and activations passed between interventions. The domain is determined by the fixed maps, operator validity, and the declared source policy. A later operator can fail at a reached state outside S_0 (Proposition 1). Empirical pools sample the closure under declared eligibility and source rules (Appendix D.1).

To compare neural and high-level outcomes, we fix a set $\mathcal{C}$ of *diagnostic continuations* before intervention. For each base, we apply valid operators from the declared family to its high-level state until no new states appear, realize this finite reachable-state fixed point, and deduplicate identical continuations. This *observable signature* groups states with the same realization. For well-defined operator action on these classes, equivalent states must agree on validity and have equivalent images under each valid operator (Remark 1, benchmark check in Appendix C.3).

The high-level target is $R_{\mathcal{G}}(z) \in \mathcal{C}$. Injecting h gives Transformer probabilities $R_\theta(c \mid h) \propto p_\theta(c \mid \text{prefix}, h)$ over $c \in \mathcal{C}$ (Appendix D.1). Endpoint scoring compares the highest-probability candidate after Φ_w with the deterministic realization $R_{\mathcal{G}}(w(z_b))$. *Compositional interchange intervention accuracy* (CIIA) averages this agreement at depth k

$$\mathrm{CIIA}_k \;=\; \Pr_{(b,w,s_{1:k})\sim\mathcal{D}_k}\Big[\;\arg\max_{c\in\mathcal{C}}\; R_\theta\big(c \mid \Phi_w(h_b; h_{s_{1:k}})\big) \;=\; R_{\mathcal{G}}\big(w(z_b)\big)\;\Big], \tag{4}$$

under the *two-stage* law $\mathcal{D}_k$, first $b \sim \mathrm{Unif}(\mathcal{B}_k)$, then $(w, s_{1:k}) \sim \mathrm{Unif}(\mathcal{W}_k(b))$. Here $\mathcal{B}_k$ contains eligible bases in the declared depth-k pool, and $\mathcal{W}_k(b)$ contains each base's eligible word/source tuples. Averaging within bases and then across bases gives each base equal weight.

Every CIIA cell also reports candidate mass $P(\mathcal{C})$, the probability assigned to the diagnostic set before renormalization. We average it with the same two-stage weights and require the mean to meet the frozen threshold $\tau = 0.5\,\bar{P}_{\mathrm{obs}}$, where $\bar{P}_{\mathrm{obs}}$ is the same mean without intervention. Cells below it fail regardless of CIIA. Per-row failures remain diagnostic. We report depths separately, paired depth loss $\mathrm{CompDrop}_k = \mathrm{CIIA}_1 - \mathrm{CIIA}_k$, and signed Δ, composed or closure-local accuracy minus unary accuracy. Appendix A.1 gives the distributional score $1 - \mathrm{JSD}$.

3.3 Complexity-bounded causal realization

With per-operator rank r_o and common nominal span m, the conservative budget $m \sum_o r_o$ counts writable directions across positions and operators. The criterion combines this budget with outcome agreement and candidate mass. Appendix A.1 gives realized per-prefix charges, subspace dimensions, and validation settings.

Definition 1 (Complexity-bounded causal realization, deterministic case)**.** Let $\mathcal{G} = (\mathcal{Z}, \mathcal{O})$ be a typed operator system, T_θ a trained Transformer with a declared common layer and nominal maximum span, $\{\Phi_o\}$ frozen rank-bounded interchange operators, and $\alpha \in [0, 1)$ a tolerance on the *average* disagreement rate. T_θ $(\alpha, r_{\max}, K)$*-realizes* $\mathcal{G}$ *up to the signature* $\mathcal{C}$ if, for every $k \le K$ and every declared pool at depth k—including pools whose base states lie in $\mathrm{Cl}_{K-1}(S_0) \setminus S_0$—the disagreement rate $1-\mathrm{CIIA}_k$ is at most α and the candidate mass satisfies $P(\mathcal{C}) \ge \tau$, under operators discovered unary-only and frozen.

Here α bounds pool-average disagreement under $\mathcal{D}_k$. Theorem 1 instead assumes one-step discrepancy at most ε at every $h \in \mathrm{Cl}_{K-1}(S_0)$ for each operator valid at the decoded state $\widehat{\pi}(h)$. Under its stated assumptions, this uniform control yields the expected composition error bound (6). Experiments report graded unary, composed, and closure-local realization profiles. Empirical support of realization requires the model and the positive control to pass, and random-model, shuffled-label, and shortcut organisms to fail, at the same α (Section 6).

4 Why closure matters for composition

4.1 A finite separation between unary and composed correctness

Consider a high-level space $Z = \{0,1\}^2$ and a low-level space $H = \{0,1\}^3$, with decoder $\widehat{\pi}(v_1, v_2, v_3) = (v_1, v_2)$. On the observational support $S_0 = \{(v_1, v_2, v_3) \in H : v_3 = v_1\}$, the third coordinate is a copy of the first. Define two high-level operators and their source-independent

low-level implementations $\Phi_a, \Phi_b : H \to H$ by

$$a(x_1, x_2) = (1, x_2), \quad \Phi_a(v_1, v_2, v_3) = (1, v_2, v_3),$$
$$b(x_1, x_2) = (x_1, x_1), \quad \Phi_b(v_1, v_2, v_3) = (v_1, v_3, v_3).$$

Proposition 1 (Unary correctness on S_0 does not imply composition)**.** *There exist finite high- and low-level systems such that every generator $o \in \mathcal{O}$ is implemented exactly on the observational support:*

$$\widehat{\pi}(\Phi_o(h)) = o(\widehat{\pi}(h)), \qquad h \in S_0,$$

while for a valid two-operator word $b \circ a$ there exists $h \in S_0$ such that

$$\widehat{\pi}(\Phi_b(\Phi_a(h))) \neq b(a(\widehat{\pi}(h)))\,.$$

Moreover, the failure occurs at an intermediate state

$$\Phi_a(h) \in \mathrm{Cl}_1(S_0) \setminus S_0.$$

On S_0, Φ_a sets the decoded first coordinate to 1 and preserves the second, as a requires. For Φ_b, $v_3 = v_1$ makes copying v_3 into the second coordinate equivalent to copying v_1. From $h_0 = (0, 0, 0)$, composition follows

$$(0, 0, 0) \xrightarrow{\Phi_a} (1, 0, 0) \xrightarrow{\Phi_b} (1, 0, 0).$$

The first step sets v_1 to 1 but leaves $v_3 = 0$, reaching $\mathrm{Cl}_1(S_0) \setminus S_0$. The second copies the stale zero into v_2, so the decoded endpoint is $(1, 0)$ instead of $b(a(0, 0)) = b(1, 0) = (1, 1)$. The witness (Appendix B.1.3, Figure 2) is exhaustively verified in Appendix B.1.4. Appendix B.2 establishes tightness of the bound below and gives counterexamples for its validity, closure, and Lipschitz requirements.

4.2 Closure-level control bounds composition error

At a reached state h, sampling an admissible source $S \sim \nu_o(\cdot \mid h)$ and applying the frozen map gives $H' = \Phi_o(h, S)$. Its law is the neural transition kernel $\widehat{K}_o(h, \cdot)$, and $\widehat{\pi}_{\#}\widehat{K}_o(h, \cdot)$ is the decoded next-state law. The W_1 distance to the point mass at $o(\widehat{\pi}(h))$, measured using d_{obs}, equals the expected one-step observable discrepancy (Lemma 2). Under the discrete metric, it is the probability of missing this target. The theorem couples neural and ideal trajectories. Each partial operator must be valid at both the decoded neural state and the ideal high-level state.

Theorem 1 (Stochastic composition on the intervention closure)**.** *Fix $K \geq 1$, $L > 0$, and $\varepsilon \geq 0$. Assume:*

(A1) ***Lipschitz high-level operators.*** *Every $o \in \mathcal{O}$ is L-Lipschitz under d_{obs}, so*

$$d_{\mathrm{obs}}(o(z), o(z')) \leq L\, d_{\mathrm{obs}}(z, z')$$

whenever $z, z' \in \mathrm{dom}(o)$.

(A2) ***Closure-level one-step control.*** *For every $h \in \mathrm{Cl}_{K-1}(S_0)$ and every $o \in \mathcal{O}$ valid at $\widehat{\pi}(h)$,*

$$W_1\Big(\widehat{\pi}_{\#}\widehat{K}_o(h, \cdot), \delta_{o(\widehat{\pi}(h))}\Big) \leq \varepsilon. \tag{5}$$

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the underlying probability space. For an intervention process of length $k \leq K$, let $(\mathcal{F}_i)_{i=0}^k$ be the history filtration generated by the initial states and the successive operator choices and state updates, defined in Appendix A.2.3.

Let $(H_i, Z_i)_{i=0}^k$ and $(O_i)_{i=1}^k$ form an admissible intervention process, meaning that, for every $i = 1, \ldots, k$,

$$O_i \in \mathcal{O}, \qquad \widehat{\pi}(H_{i-1}) \in \mathrm{dom}(O_i), \qquad Z_{i-1} \in \mathrm{dom}(O_i)$$

almost surely. Each O_i is $\mathcal{F}_{i-1}$-measurable,

$$H_0 \in S_0, \qquad Z_0 = \widehat{\pi}(H_0)$$

almost surely, and, for $i = 1, \ldots, k$,

$$\mathcal{L}(H_i \mid \mathcal{F}_{i-1}) = \widehat{K}_{O_i}(H_{i-1}, \cdot), \qquad Z_i = O_i(Z_{i-1})$$

almost surely.

Then

$$\mathbb{E}[d_{\mathrm{obs}}(\widehat{\pi}(H_k), Z_k)] \le \varepsilon \sum_{j=0}^{k-1} L^j. \tag{6}$$

Consequently,

$$W_1(\mathcal{L}(\widehat{\pi}(H_k)), \mathcal{L}(Z_k)) \le \varepsilon \sum_{j=0}^{k-1} L^j. \tag{7}$$

In particular, if $L = 1$, then

$$W_1(\mathcal{L}(\widehat{\pi}(H_k)), \mathcal{L}(Z_k)) \le k\varepsilon.$$

For $i = 0, \ldots, k$, let $e_i = \mathbb{E}[d_{\mathrm{obs}}(\widehat{\pi}(H_i), Z_i)]$. Appendix B.1.1 derives

$$e_i \le \varepsilon + L e_{i-1}, \qquad i = 1, \ldots, k.$$

Closure-level control bounds the expected discrepancy of the decoded neural update from $O_i(\widehat{\pi}(H_{i-1}))$ by ε, the new error at step i. Applying the L-Lipschitz operator O_i to $\widehat{\pi}(H_{i-1})$ and Z_{i-1} propagates their discrepancy by at most Le_{i-1} after averaging. Before step i, H_{i-1} lies in $\mathrm{Cl}_{i-1}(S_0) \subseteq \mathrm{Cl}_{K-1}(S_0)$ almost surely, so one-step control applies throughout. Starting from $Z_0 = \widehat{\pi}(H_0)$ gives $e_0 = 0$. Iteration yields (6), or $k\varepsilon$ when $L = 1$.

4.3 From the bound to endpoint and local measurements

For a base with generator state z_b, two frozen interventions produce the trajectory $h_0 \to h_1 \to h_2$, starting from its un-intervened neural state h_0. We evaluate endpoint and closure-local accuracy by comparing the same decoded endpoint with two targets in the observable state space

$$\begin{aligned} \text{endpoint} \quad & \widehat{\pi}(h_2) \quad \text{versus} \quad o_2(o_1(z_b)), \\ \text{closure-local} \quad & \widehat{\pi}(h_2) \quad \text{versus} \quad o_2(\widehat{\pi}(h_1)). \end{aligned}$$

Endpoint accuracy tests the complete sequence from z_b. Closure-local accuracy tests the action of o_2 from the decoded intermediate state, including eligible rows where the model decodes the first-step outcome incorrectly. Local scoring requires o_2 to be valid at $\widehat{\pi}(h_1)$ and its target to be measurable in the declared signature. If $\widehat{\pi}(h_1) = o_1(z_b)$, the two targets coincide.

Corollary 1 covers deterministic interventions. The theorem starts its reference trajectory at $\widehat{\pi}(H_0)$, while endpoint scoring starts at the generator state z_b. With $Z_0 = z_b$ and the remaining assumptions, including validity along both trajectories, Remark 2 gives

$$\mathbb{E}[d_{\mathrm{obs}}(\widehat{\pi}(H_k), Z_k)] \le L^k e_0 + \varepsilon \sum_{j=0}^{k-1} L^j,$$

where $e_0 = \mathbb{E}[d_{\mathrm{obs}}(\widehat{\pi}(H_0), z_b)]$ is the base decoding error (measurements on the replication models in Table 12). Under the discrete metric, the expectation is the endpoint error probability. Averaging this disagreement under $\mathcal{D}_k$ gives $1 - \mathrm{CIIA}_k$. Experiments estimate endpoint and closure-local accuracy from the same interventions and prescribed source policy on finite pools (Section 5).

5 OpClosure and experimental design

5.1 Why OpClosure provides a test of reusable operators

OpClosure's deterministic generator provides high-level states, exact counterfactual targets, and frame-matched donors for endpoint and local evaluation (Section 4.3, Figure 3). Validity rules determine executable sequences. Separate pools test IID, lexical and composition holdouts, order, depth, intervention closure, and surface perturbations. Appendix C.3 gives the records, split rules, and passing leakage checks. The generator places grammatical cues before the intervention site and an engineered, English-like echo after it (Appendix C.3.2). In *the chefs did not eat the cake , and the cake is sweet , did they ?*, the first comma marks the site. A matched present-tense donor uses

do not eat before that comma, with other state fields unchanged. We apply the learned interchange over the span ending at that comma. The target continuation keeps the theme-property clause and ends in *do they ?*. Scoring tests whether the model makes this change.

Two TENSE applications return to the original observable state, giving length two and effective depth zero. Effective depth is the shortest valid word length reaching the same observable endpoint from the same base (Definition 2). Before sampling, OpClosure-v2 retains words whose effective depth equals their length, including in the closure pool. Genuine depth-3 endpoints thus require three operations (Table 1).

The seven-family inventory has six autoregressive candidates. WH is frame-only because object questions lack an echo-tail site. The five active operators in the primary instance are NUM, TENSE, POL, BIND, and PASS. The validation candidate-mass gate excludes CM there and admits it in instance 90904 (Table 13).

5.2 MODELS, FITTING, AND EVALUATION

We fixed settings on the OpClosure development suite before generating the independent confirmatory OpClosure-v2 instance (Table 7). The six prespecified contrasts compare closure-local accuracy (P1) and genuine depth-3 accuracy (P2) with unary accuracy. On IID pairs, P3 and P4 compare standard models with random and shuffled controls, respectively, and P5 compares shortcut with standard models. P6 compares held with seen signatures (Table 10). Table 2 summarizes development-suite counts and model coverage, with development-suite comparisons in Appendix E.1.

Model training uses the full lexicon. Map fitting uses class-A unary rows with model weights fixed, so the class-B holdout tests intervention transfer. Validation and freezing follow Section 3.1. The standard model learns next-token prediction from the observational corpus. Its matched random initialization tests behavior before training. The shortcut model omits present-tense negative examples under the same training budget, inducing PRES $\Rightarrow$ POS. Whole-label shuffling holds standard weights fixed and permutes operator labels without fixed points. Each fitting slot receives its assigned operator's complete unary rows, matched sources, and targets (Appendices D.2 and D.1).

Unary interchange intervention training (Geiger et al., 2022) creates the modular reference by supervising source-to-base copies of designated coordinate blocks with unary counterfactual targets. Model training with composed interchanges adds correct or stale composed targets (Section 6.3). The direct arm of the map-family comparison instead fits maps on depth-2 validation rows with model weights fixed. This comparison also reports probing/unary DAS (Geiger et al., 2024) and unrestricted alignments under a prospective protocol with exploratory analysis status (Table 30). Descriptive controls assess sensitivity to surface form, intervention norm and site, and off-target effects (Appendix D.3).

Synthetic readout teacher-forces the base prefix and renormalizes full-continuation probabilities (Section 3.2). The endpoint measure is held-out CIIA_3. Paired synthetic contrasts average eligible rows and model-training seeds within each base, then average differences over shared bases. The six frozen v2 contrasts use 95% paired base-bootstrap intervals, sign-flip tests, and Benjamini–Hochberg correction. Confirmation requires the base-bootstrap and original hierarchical estimates to agree under the declared rule (Table 10). Model-training seeds vary weights, operator-fitting seeds refit maps on fixed weights, and benchmark instances vary the dataset. Appendix E.3 gives supplementary resampling procedures.

5.3 NUMBER AND TENSE IN PRETRAINED LANGUAGE MODELS

Family A tests number and tense interchanges on controlled English continuations such as *eats/eat* and *is/are/was/were*. Donors differ before the site in subject number or a tense adverb, with other grammatical variables fixed. A tokenizer-specific gate checks the prefix difference. We truncate candidates to the shortest common token length that distinguishes them pairwise, separating passive forms at the auxiliary and active-present forms at the verb.

With model weights fixed, we fit unary maps using three operator-fitting seeds for each of seven open-weight base models (Section 6.4). Seed 0 selects a common layer and nominal span and per-operator ranks on validation frames. Seeds 1 and 2 reuse this configuration. Frozen maps are tested

in both depth-2 TENSE×NUM composition orders on two lexical frames held out from fitting and selection (Appendix I.2.1).

Two independent scorers and one adjudicator produced the final round-3 Family-A inclusion mask, retaining all targets. Each eligible high-level step must change the scored candidate. Active-past number is undefined because both numbers share forms such as *served*, so rows whose number step lands there are excluded. Family-B caused-motion stimuli (Goldberg, 1995) and full premise-hypothesis annotations appear in Appendix I, with naturalness, plausibility, and item licensing reported as exploratory.

6 EMPIRICAL CHARACTERIZATION OF OPERATOR REUSE

6.1 FROZEN OPERATORS LOSE ACCURACY UNDER COMPOSITION AND ON REACHED STATES

On the OpClosure-v2 confirmatory instance, the genuine depth-3 minus unary accuracy contrast is -0.387 (95% base-bootstrap CI [-0.460, -0.313]), with the same direction in all five independently trained models. These endpoints require three operations (Definition 2, Table 1). Of the six preregistered contrasts, 4 are confirmed under both estimators (P1, P2, P3, P4, Table 10). On IID pairs, standard models exceed random and shuffled controls by +0.391 and +0.418, respectively.

An independent-per-step reference (u^3, with $u = 0.775$ on the same 85 bases) attributes about 80% of the drop to ordinary accumulation and leaves a composition-specific excess of -0.078 (Appendix F.4).

For the five original confirmatory models, the pooled closure-local minus unary contrast is -0.242 (95% base-bootstrap CI [-0.280, -0.205]) on 277 shared bases, negative on all 5 model seeds (Table 11).

A descriptive analysis of five cross-machine replication organisms trained under the frozen configuration also matches the second operator within each base sentence. Its strict closure-local minus unary contrast is -0.221, compared with -0.220 when restricted to correctly decoded first steps. The later operator thus loses local accuracy even after a correct first step. The deficit also persists against same-operator observational sentences matching the reached state's structural profile (Table 12). Both the pooled closure-local and genuine depth-3 contrasts are confirmed in each of three further benchmark instances. On their confirmatory models, strict same-base, same-operator gaps lie between -0.170 and -0.258 (Table 13).

In the terms of Definition 1, unary validation admits the standard models at a median α^* of 0.10 to 0.22 on every instance; the full criterion places them at 0.63 to 0.88 (Table 9).

6.2 CONTROLLED REFERENCES CALIBRATE THE CLOSURE-LOCAL DEFICIT

We calibrate the measured closure-local deficit against a unary-trained modular reference. Its designated-block mechanism composes exactly under its own interventions. We evaluate the reference at its designated site using both these oracle maps and the protocol's fitted maps. Across the replication set and three further instances, its mean strict closure-local minus unary gaps are -0.156 and -0.157, respectively. On the same model, a constructed stale-copy readout restores every other designated block to its pre-intervention value after each operator, adapting Proposition 1. Under fitted maps, its mean gap is -0.861, while unary accuracy remains within 0.02 of the reference under the same maps on every set (Table 16).

The standard models' mean strict gap under fitted maps is -0.213, a paired excess of -0.056 over the reference. Most of the reference's own gap is donor geometry: on the replication models and two further instances, aligning the second donor by grammatical role, with everything else frozen, shrinks the reference's gap to -0.048 while the standard models keep -0.124 [-0.145, -0.105]. The paired model-attributable excess is then -0.076 [-0.105, -0.047] (-0.089 against exact interchanges), negative on every set (Table 17).

6.3 CLOSURE TRAINING REMOVES THE DEFICIT AND STALE TRAINING REPRODUCES IT

Adding correct composed interchange targets to model training raises closure-local accuracy from 0.830 for the unary-trained modular reference to 0.942 for the model trained with correct composed targets on the same replication dataset under fitted maps (Table 18). Training on reached states closes most of the gap, shifting the strict closure-local minus unary contrast from -0.177 to -0.087 (95% paired base-bootstrap CI [-0.132, -0.045]).

A second family member restores NUM to its base value in the composed training target. It retains unary accuracy 0.959 and loses closure-local accuracy exactly where it was taught to, scoring 0.769 when the first or second operator is NUM and 0.952 elsewhere.

Across all six members, the fraction of measurable closure rows involving a stale variable correlates with closure-local loss, with descriptive Spearman +0.90 under fitted maps and +0.99 under oracle maps.

6.4 RANK-BOUNDED INTERCHANGES REALIZE NUMBER AND TENSE COMPOSITION

We test frozen number and tense interchanges in Pythia-1.4B, Pythia-2.8B, Pythia-410M, Qwen2.5-1.5B, SmolLM2-1.7B, SmolLM2-360M, SmolLM3-3B, spanning three families and 360M to 3B parameters. Under the completed number–tense protocol, both composition orders are evaluated on held-out lexical frames with the final round-3 inclusion mask. All seven models pass the capability and candidate-mass gates (Table 23).

SmolLM2-360M and SmolLM3-3B sit at ceiling on every evaluation pool of the number–tense protocol. Six models attain mean composed CIIA of at least 0.889 in both orders; Pythia-2.8B's composed means (0.667, 0.639) include one seed whose rank-8 NUM map does not converge (Appendix I.2.1). These results demonstrate that a rank-bounded interchange at one site realizes both operators and their compositions across three families and sizes from 360M to 3B.

On the frozen test frames, the item-level decomposition gives pooled local-minus-unary gaps from -0.222 to $+0.000$ over 8 to 9 items per model (Table 24).

The copula agreement test scores the same pair at scale on passive items whose copula carries both number and tense, 1,220 per word on 305 test nouns never used for fitting or selection. The closure-local loss is confirmed in five of the six models that pass its reproduction gate, up to 0.127 per model; pooled over them, Δ is -0.047 $[-0.057, -0.038]$ for TENSE after NUM and -0.038 $[-0.046, -0.031]$ for NUM after TENSE, and 1,872 of 109,636 steps read the first variable's stale cue. Pythia-6.9B gives $+0.004$ and -0.008 (Appendix I.3).

7 DISCUSSION

Unary-only fits remain effective on pairs, triples, depth four and closure states, and their uneven local losses make reached states part of a causal account of reuse. The unary-trained modular reference calibrates these losses and identifies the model-attributable excess under ordinary training; correct composed targets close most of the gap, while stale targets produce the specified local failures. In controlled English, rank-bounded interchanges realize number and tense and their compositions across three pretrained families, and the closure-local loss reappears there on held-out copula items. Evaluating endpoints and local fidelity together tests both the intended transformation and each operator's reliability after earlier interventions.

8 CONCLUSION

Complexity-bounded causal realization tests whether fixed linguistic interventions compose on the states they generate. A finite counterexample separates unary from composed correctness, and uniform closure-level control yields a tight stochastic composition bound under Lipschitz and validity conditions, with per-assumption counterexamples. OpClosure provides exact counterfactuals for measuring reuse losses and calibrating their model-attributable component, and composed model training closes most of the local gap. In pretrained models, rank-bounded interchanges realize number and tense composition, and the closure-local loss reappears on held-out items.

REPRODUCIBILITY STATEMENT

The appendices provide mathematical proofs. Exhaustive enumeration verifies the fixed finite counterexample of Proposition 1 (`verify_counterexample.py`) and the minimality analysis (`verify_review_claims.py`). Numerical tests check the bound of Corollary 1 on random systems and at its tightness point. They also evaluate explicit counterexamples obtained by removing the Lipschitz condition, closure-level one-step control, or validity at the decoded state (`verify_metric_definitions.py`, `verify_review_claims.py`). The stochastic theorem, its two supporting lemmas, and Proposition 3 are established by the mathematical proofs in the appendices.

The OpClosure benchmark is generated deterministically from a fixed seed (`scripts/generate_synthetic_grammar.py`), and a validator checks its split invariants (`scripts/validate_splits.py`). The intervention pipeline, metric aggregation, and the frozen statistical procedure (paired bootstrap over base examples, Benjamini–Hochberg correction over each suite's own frozen family: seven members for the OpClosure development suite, six for the OpClosure-v2 confirmatory suite) are released as code with the exact seeds. Admissible result files must pass the strict version-3 schema, finite-number, dataset/selection provenance, and complete-suite checks (`results/result_schema.json`; `scripts/validate_results.py`). An anonymized repository containing all code, data, and audit ledgers accompanies the submission.

AI USE STATEMENT

Generative AI systems were used for manuscript drafting, editing and generating synthetic datasets. The released scripts perform the finite enumeration and numerical tests described in the reproducibility statement. The authors remain responsible for all claims, proofs, code, data, and artifacts.

REFERENCES

Nura Aljaafari, Danilo Carvalho, and Andre Freitas. Where do LLMs compose meaning? a layerwise analysis of compositional robustness. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4622–4646, 2026. doi: 10.18653/v1/2026.eacl-long.214. URL https://aclanthology.org/2026.eacl-long.214/.

Zhiyu An and Wan Du. Representational homomorphism predicts and improves compositional generalization in transformer language model. *arXiv preprint arXiv:2601.18858*, 2026. URL https://arxiv.org/abs/2601.18858.

Aryaman Arora, Dan Jurafsky, and Christopher Potts. CausalGym: Benchmarking causal interpretability methods on linguistic tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14638–14663, 2024.

Sander Beckers, Frederick Eberhardt, and Joseph Y. Halpern. Approximate causal abstractions. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*, volume 115 of *Proceedings of Machine Learning Research*, pp. 606–615, 2020.

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*, 2025.

Lawrence Chan, Adrià Garriga-Alonso, Nicholas Goldowsky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. Causal scrubbing: A method for rigorously testing interpretability hypotheses. AI Alignment Forum, 2022. URL https://www.alignmentforum.org/posts/JvZhhzycHu2Yd57RN/causal-scrubbing-a-method-for-rigorously-testing.

Sonakshi Chauhan, Maheep Chaudhary, Choy Kwan Kiu, Samuel Nellessen, and Nandi Schoots. Punctuations and predicates in language models. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 5622–5636, 2026. doi: 10.18653/v1/2026.findings-eacl.297. URL https://aclanthology.org/2026.findings-eacl.297/.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems 36*, 2023.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021.

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

Matthew Finlayson, Aaron Mueller, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, and Yonatan Belinkov. Causal analysis of syntactic agreement mechanisms in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021.

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

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. Inducing causal structure for interpretable neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 7324–7338, 2022.

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

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

Adele E. Goldberg. *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press, 1995.

Rhys Gould, Euan Ong, George Ogden, and Arthur Conmy. Successor heads: Recurring, interpretable attention heads in the wild. In *International Conference on Learning Representations (ICLR)*, 2024.

Satchel Grant, Simon Jerome Han, Alexa R. Tartaglini, and Christopher Potts. Addressing divergent representations from causal interventions on neural networks. *arXiv preprint arXiv:2511.04638*, 2025.

Itamar Hadad, Guy Katz, and Shahaf Bassan. Formal mechanistic interpretability: Automated circuit discovery with provable guarantees. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Timsb74vIY.

Michael Hahn. Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171, 2020.

Irene Heim. *The Semantics of Definite and Indefinite Noun Phrases*. PhD thesis, University of Massachusetts Amherst, 1982. URL https://semanticsarchive.net/Archive/jA2YTJmN/Heim%20Dissertation%20with%20Hyperlinks.pdf.

Jing Huang, Zhengxuan Wu, Kyle Mahowald, and Christopher Potts. Inducing character-level structure in subword-based language models with type-level interchange intervention training. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 12163–12180, 2023. doi: 10.18653/v1/2023.findings-acl.770. URL https://aclanthology.org/2023.findings-acl.770/.

Jing Huang, Zhengxuan Wu, Christopher Potts, Mor Geva, and Atticus Geiger. RAVEL: Evaluating interpretability methods on disentangling language model representations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8669–8687, 2024.

Apoorv Khandelwal and Ellie Pavlick. How do language models compose functions? *arXiv preprint arXiv:2510.01685*, 2026. URL https://arxiv.org/abs/2510.01685. First posted 2025; revised 2026.

Alina Klerings, Jannik Brinkmann, Daniel Ruffinelli, and Simone Paolo Ponzetto. Steering language models in multi-token generation: A case study on tense and aspect. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025.

Arinbjorn Kolbeinsson, Kyle O'Brien, Tianjin Huang, Shanghua Gao, Shiwei Liu, Jonathan Richard Schwarz, Anurag Vaidya, Faisal Mahmood, Marinka Zitnik, Tianlong Chen, and Thomas Hartvigsen. Composable interventions for language models. In *International Conference on Learning Representations*, 2025.

Joachim Lambek. The mathematics of sentence structure. *The American Mathematical Monthly*, 65 (3):154–170, 1958.

Michael A. Lepori, Thomas Serre, and Ellie Pavlick. Break it down: Evidence for structural compositionality in neural networks. In *Advances in Neural Information Processing Systems 36*, 2023.

Baohao Liao and Christof Monz. 3-in-1: 2d rotary adaptation for efficient finetuning, efficient batching and composability. In *Advances in Neural Information Processing Systems 37*, 2024. doi: 10.52202/079017-1104. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/3dbcadb7beedc2afe32bb23f75dd30ec-Abstract-Conference.html.

David Lindner, János Kramár, Sebastian Farquhar, Matthew Rahtz, Tom McGrath, and Vladimir Mikulik. Tracr: Compiled transformers as a laboratory for interpretability. In *Advances in Neural Information Processing Systems 36*, 2023.

Robin Lorenz and Sean Tull. Causal and compositional abstraction. *arXiv preprint arXiv:2602.16612*, 2026.

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

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

Maxime Méloux, Tiago Pimentel, François Portet, and Maxime Peyrard. Validating causal abstraction metrics on simulated complex systems. *arXiv preprint arXiv:2607.00267*, 2026.

Philipp Mondorf, Sondre Wold, and Barbara Plank. Circuit compositions: Exploring modular structures in transformer-based language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14934–14955, 2025.

Richard Montague. Universal grammar. *Theoria*, 36(3):373–398, 1970.

Nils Palumbo, Ravi Mangal, Zifan Wang, Saranya Vijayakumar, Corina S. Pasareanu, and Somesh Jha. Validating mechanistic interpretations: An axiomatic approach. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 47509–47544, 2025. URL https://proceedings.mlr.press/v267/palumbo25a.html.

Eigil F. Rischel and Sebastian Weichwald. Compositional abstraction error and a category of causal models. In *Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pp. 1013–1023, 2021.

Aryan Sharma, Cutter Dawes, and Shivam Raval. Dissociating decodability and causal use in bracket-sequence transformers. *arXiv preprint arXiv:2604.22128*, 2026. URL https://arxiv.org/abs/2604.22128.

Paul Smolensky. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, 46(1–2):159–216, 1990.

Lena Strobl, William Merrill, Gail Weiss, David Chiang, and Dana Angluin. What formal languages can transformers express? A survey. *Transactions of the Association for Computational Linguistics*, 12:543–561, 2024.

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 38*, 2025.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 407–416, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.25. URL https://aclanthology.org/2024.blackboxnlp-1.25/.

Cheng Tang, Brenden Lake, and Mehrdad Jazayeri. Circuit explained: How does a transformer perform compositional generalization. *PLOS ONE*, 21(2):e0340088, 2026. doi: 10.1371/journal.pone.0340088. URL https://doi.org/10.1371/journal.pone.0340088.

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.

Gail Weiss, Yoav Goldberg, and Eran Yahav. Thinking like transformers. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11080–11090, 2021.

Zhengxuan Wu, Atticus Geiger, Aryaman Arora, Jing Huang, Zheng Wang, Noah Goodman, Christopher D. Manning, and Christopher Potts. pyvene: A library for understanding and improving PyTorch models via interventions. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations)*, pp. 158–165, 2024. doi: 10.18653/v1/2024.naacl-demo.16. URL https://aclanthology.org/2024.naacl-demo.16/.

Xining Xun. Causal structure is inducible but functionally decoupled: The routing/readout boundary of a typed mechanism library. *arXiv preprint arXiv:2608.11767*, August 2026. URL https://arxiv.org/abs/2608.11767. Submitted 12 August 2026.

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 Formal objects and observable semantics

The definitions in Appendix A.1 specify the typed states, intervention budgets, and observable semantics used by the criterion in Section 3. Appendix A.2 develops the source kernels and intervention closure, then defines the one-step error and admissible processes used in Theorem 1.

A.1 Complete definitions

This appendix states the general formulation of which §3 is the deterministic, tied-projector span special case. Notation is unchanged.

A.1.1 TYPED STATES, GENERAL FORM

A *linguistic state* may be taken as $z_t = (g_t, \sigma_t, \omega_t)$, where g_t is a finite typed relational structure over entities, events, grammatical functions, semantic roles, and construction instances; σ_t is a finite multiset of open obligations (unresolved agreement, filler–gap, attachment, scope, or coreference constraints); and ω_t is a discourse state. The benchmark of this paper operationalizes: entities and role bindings (a fragment of g_t), agreement and filler–gap obligations (a fragment of σ_t; per-token generator metadata are released with the data), and no discourse component (ω_t trivial). The formal results concern the operationalized fragment of this state description.

A.1.2 STOCHASTIC OPERATORS AND KERNELS

A stochastic generator would be a (possibly partial) Markov kernel $o : \mathrm{dom}(o) \subseteq \mathcal{Z}_\tau \to \Delta(\mathcal{Z}_{\tau'})$, of which deterministic operators are the point-mass case. The high-level operators used in this paper are deterministic. In Theorem 1, randomness enters through low-level source selection by ν_o, inducing the transition kernels $\widehat{K}_o$ of Appendix A.2. The theorem covers validity-preserving intervention processes.

A.1.3 SITES AND MULTI-SITE INTERVENTIONS

A *site set* $\mathcal{S} = \{(\ell_1, t_1), \ldots, (\ell_q, t_q)\}$ selects residual-stream positions; the joint state is the product $h = (h^{(1)}, \ldots, h^{(q)}) \in \prod_i \mathbb{R}^d$. A block interchange operator applies (1) per site. In the general case each site may have its own $U_o^{(i)}$, and the complexity account charges $\sum_i r_o^{(i)}$ and $|\mathcal{S}|$. The implemented protocol ties one U_o across an input-dependent right-aligned block with nominal maximum $m \in \{1, 4, 8, 12\}$: for a prefix whose site is t^*, the actual block is $\mathcal{S}_{\le m}(x) = \{t^* - k : 0 \le k < q_x\}$, where $q_x = \min\{m, t^* + 1\}$ excludes positions left of the first ordinary token. Base and source blocks are right-aligned by offset from their own endpoints. If a source has $q_s < q_b$ available offsets, its leftmost activation is edge-replicated for base offsets $k \ge q_s$; source offsets beyond q_b are unused. Equation (1) acts coordinatewise on this lifted base/source block in $(\mathbb{R}^d)^{q_b}$, with one projector tied across offsets. We report both the unique-subspace dimension $r_o(d - r_o)$ and the conservative rank–site budget $r_o m$; the realized per-prefix charge is $r_o q_x$.

Across the operator inventory, the conservative budget is $m \sum_o r_o$, its realized charge at a prefix is $q_x \sum_o r_o$, and the Grassmannian dimension is $\sum_o r_o(d - r_o)$. Validation selection and its recorded layer/span/rank grid are described in Appendices D.2 and D.1. Table 29 reports a separate exploratory test-set rank sweep after validation selection was frozen. The classical description-length bound is stated in Proposition 2.

A.1.4 SAMPLING DISTRIBUTIONS

Empirical pool averages use the declared two-stage sampling law. $\mathcal{D}_k$ first draws a base uniformly, then draws a valid (word, sources) tuple uniformly within that base in the named evaluation pool at depth k (pool construction: Appendix C.3; source matching: Appendix D.1). Where pools are capped for compute, the released, model-independent plan uses independently seeded base order and word-within-base order per pool, traversed round-robin before the cap.

A.1.5 OUTCOME DISCREPANCIES

For distributions p, q on the finite candidate set $\mathcal{C}$, $\mathrm{JSD}(p, q) = \frac{1}{2}\mathrm{KL}(p \,\|\, m) + \frac{1}{2}\mathrm{KL}(q \,\|\, m)$ with $m = \frac{1}{2}(p + q)$ and logarithms base 2, so $\mathrm{JSD} \in [0, 1]$, symmetric, and zero iff $p = q$ (verified numerically in `verify_metric_definitions.py`). CIIA_k is the indicator-average (4); in the deterministic case it equals one minus the disagreement rate of the argmax decision rule. $\mathrm{CompDrop}_k = \mathrm{CIIA}_1 - \mathrm{CIIA}_k$ is computed on paired base examples; because both terms are accuracies of the same model at the same site, the subtraction is meaningful as a within-model degradation, and we report it alongside both component accuracies.

A.1.6 DETERMINISTIC VS. STOCHASTIC SEMANTICS

The high-level reference $R_{\mathcal{G}}(z)$ is deterministic throughout (the generator's realization function). The low-level outcome $R_\theta(\cdot \mid h)$ is a renormalized conditional distribution. CIIA compares the argmax of the latter to the former; JSD compares the distribution to the point mass.

Because candidate-set sizes differ, the chance-normalized depth score is $(\mathrm{CIIA}_k - c_k)/(1 - c_k)$, where c_k is the mean reciprocal candidate-set size. The development suite reports raw paired $\mathrm{CompDrop}_k$. Its row and aggregate artifacts omit c_k, which the normalized score requires.

A.1.7 OBSERVABLE QUOTIENT AND IDENTIFIABILITY

Remark 1 (Observable quotient). By the type-separation assumption of Appendix A.2, the diagnostic set $\mathcal{C}$ induces

$$z \sim z' \quad \Longleftrightarrow \quad R_{\mathcal{G}}(z) = R_{\mathcal{G}}(z').$$

Hence Definition 1 constrains the action only on the observable quotient $\mathcal{Z}/\sim$.

For the quotient action of an operator o to be well defined, observational equivalence must be respected both by its validity domain and by its action. This congruence property is assumed in Theorem 1. The benchmark, however, does not assume congruence a priori; instead, it audits the property empirically and reports violating pairs (Appendix C.3). Such violations fall outside the theorem's guarantee and may allow observationally indistinguishable states to diverge under later interventions.

Multiple underlying mechanisms can satisfy the same realization criterion relative to $\mathcal{C}$.

A.2 SETTING: THE OBJECTS THE RESULTS QUANTIFY OVER

This appendix fixes, once, the objects that Theorem 1, Corollary 1, and Definition 2 quantify over. The empirical protocol of §3 is the deterministic special case noted below; its interchange operator (1) is the map Φ_o defined here. Proposition 1 uses the separate finite construction given in Appendix B.1.3.

A.2.1 SETUP

Let $\mathcal{T}$ be a finite set of state types. For each $\tau \in \mathcal{T}$, let Z_τ be a finite set of high-level states of type τ, and define the typed state space

$$Z := \bigsqcup_{\tau \in \mathcal{T}} Z_\tau.$$

Thus every state $z \in Z$ belongs to a unique type τ.

Let $\mathcal{O}$ be a finite collection of typed generator operators. Each operator $o \in \mathcal{O}$ is a partial map

$$o : \mathrm{dom}(o) \subseteq Z_\tau \longrightarrow Z_{\tau'}$$

for some source type τ and target type τ' depending on o. The set

$$\mathrm{dom}(o) \subseteq Z_\tau$$

is the validity domain of o: the operation is defined only for states in $\mathrm{dom}(o)$.

We write

$$G = (Z, \mathcal{O})$$

for the resulting typed high-level operator system, consisting of the high-level state space Z together with the family $\mathcal{O}$ of valid generator transformations acting on it.

A length-k operator word is an ordered composition of a type-compatible sequence of generators

$$o_i : \mathrm{dom}(o_i) \subseteq Z_{\tau_{i-1}} \longrightarrow Z_{\tau_i}, \qquad i = 1, \ldots, k.$$

The sequence therefore has the type structure

$$Z_{\tau_0} \xrightarrow{o_1} Z_{\tau_1} \xrightarrow{o_2} \cdots \xrightarrow{o_k} Z_{\tau_k}.$$

Its composition induces a partial map

$$w := o_k \circ \cdots \circ o_1 : \mathrm{dom}(w) \subseteq Z_{\tau_0} \longrightarrow Z_{\tau_k},$$

where

$$\mathrm{dom}(w) := \{z \in Z_{\tau_0} : (o_{i-1} \circ \cdots \circ o_1)(z) \in \mathrm{dom}(o_i) \text{ for every } i = 1, \ldots, k\},$$

with the convention that, for $i = 1$,

$$(o_{i-1} \circ \cdots \circ o_1)(z) = z.$$

Equivalently, for an initial state $z \in Z_{\tau_0}$, define recursively

$$z_0 := z, \qquad z_i := o_i(z_{i-1}), \qquad i = 1, \ldots, k,$$

whenever the right-hand side is defined. Then

$$z \in \mathrm{dom}(w)$$

if and only if

$$z_{i-1} \in \mathrm{dom}(o_i) \qquad \text{for every } i = 1, \ldots, k.$$

When this condition holds, we say that w is *valid at* z, and its value is

$$w(z) = z_k = o_k\big(o_{k-1}(\cdots o_1(z) \cdots)\big).$$

We denote by

$$\mathcal{W}_k(z)$$

the set of length-k operator words whose induced partial maps are defined at z, equivalently, the set of valid length-k words at z. We further write

$$\mathcal{W}_{\le K}(z) := \bigcup_{k=0}^{K} \mathcal{W}_k(z)$$

for the set of valid words of length at most K, where the empty word is interpreted as the identity map on the type containing z.

Operator order matters: the generators need not commute, and validity may itself depend on the order in which they are applied. Thus $o_2 \circ o_1$ and $o_1 \circ o_2$ may differ in value, or one may be valid at a state z while the other is not.

Let $\mathcal{C}$ be a finite, predeclared set of diagnostic continuations. For each high-level state $z \in Z$, let

$$R_G : Z \longrightarrow \mathcal{C}$$

assign the continuation representing the observable realization of z.

We assume that diagnostic realizations separate state types:

$$R_G(z) = R_G(z') \quad \Longrightarrow \quad \tau(z) = \tau(z').$$

Distinct high-level states of the same type may have the same diagnostic realization. Since each $z \in Z$ belongs to a unique type, write $\tau(z) \in \mathcal{T}$ for its type, and define

$$z \sim_{\mathcal{C}} z' \quad \Longleftrightarrow \quad \tau(z) = \tau(z') \quad \text{and} \quad R_G(z) = R_G(z').$$

The corresponding observable state space is

$$\overline{Z} := Z/\!\sim_{\mathcal{C}}.$$

Thus $\overline{Z}$ retains exactly the distinctions between high-level states that can be resolved by the diagnostic continuation set.

Let H be the neural intervention state space. Given $h \in H$, the model assigns a score $R_\theta(c \mid h)$ to each $c \in \mathcal{C}$.

Let

$$\mathcal{C}_G := R_G(Z) \subseteq \mathcal{C}$$

be the set of diagnostic continuations realized by high-level states. Fix a deterministic total ordering $\prec$ on $\mathcal{C}_G$, used only to break ties. For each $h \in H$, define

$$c^*(h) := \min_{\prec} \arg\max_{c \in \mathcal{C}_G} R_\theta(c \mid h).$$

Equip $\overline{Z}$ with the discrete σ-algebra, and assume that $h \mapsto R_\theta(c \mid h)$ is Borel measurable for every $c \in \mathcal{C}_G$. Since $\mathcal{C}_G$ is finite and ties are broken deterministically, the map $h \mapsto c^*(h)$ is measurable.

Then the observable decoder

$$\widehat{\pi} : H \longrightarrow \overline{Z}$$

is defined by

$$\widehat{\pi}(h) := [z]_{\mathcal{C}} \qquad \text{for any } z \in Z \text{ such that } R_G(z) = c^*(h).$$

Such a state z exists because $c^*(h) \in \mathcal{C}_G = R_G(Z)$. By the type-separation assumption and the definition of $\sim_{\mathcal{C}}$, the equivalence class $[z]_{\mathcal{C}}$ is independent of the choice of z. Hence $\widehat{\pi}$ is well defined and measurable.

Empirical intervention accuracy compares states in $\overline{Z}$, the observable quotient of Z.

For Theorem 1 and Corollary 1, we assume that each high-level operator respects observational equivalence, including its validity domain: for every $o \in \mathcal{O}$,

$$z \sim_{\mathcal{C}} z' \quad \Longrightarrow \quad (z \in \mathrm{dom}(o) \iff z' \in \mathrm{dom}(o)),$$

and whenever these equivalent conditions hold,

$$o(z) \sim_{\mathcal{C}} o(z').$$

Hence, under this assumption, each generator induces a well-defined partial map on $\overline{Z}$, which we continue to denote by o.

Let d denote the residual-stream dimension. For each operator $o \in \mathcal{O}$, let

$$h_b, h_s \in \mathbb{R}^d$$

denote respectively the base activation and an admissible source activation at the intervention site. The source h_s is taken from a matched example whose high-level state differs from the base in the variable written by o.

For each operator $o \in \mathcal{O}$, let

$$U_o \in \mathbb{R}^{d \times r_o}, \qquad U_o^\top U_o = I_{r_o}, \qquad 1 \le r_o \le r_{\max},$$

where r_o is the intervention rank assigned to o, $r_{\max}$ is the declared maximum rank, and I_{r_o} is the $r_o \times r_o$ identity matrix. The columns of U_o form an orthonormal basis for an r_o-dimensional intervention subspace of $\mathbb{R}^d$.

Define the orthogonal projector onto this subspace by

$$P_o := U_o U_o^\top \in \mathbb{R}^{d \times d}.$$

For base and source activations

$$h_b, h_s \in \mathbb{R}^d,$$

the implemented low-level intervention is

$$\Phi_o(h_b, h_s) := (I_d - P_o)h_b + P_o h_s,$$

where I_d is the $d \times d$ identity matrix. Equivalently,

$$\Phi_o(h_b, h_s) = h_b + P_o(h_s - h_b).$$

Thus the component of h_b orthogonal to the learned o-subspace is preserved, while its component inside that subspace is replaced by the corresponding component of h_s.

Let

$$(H, \mathcal{H}) := (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$$

be the measurable neural intervention state space, where $\mathcal{B}(\mathbb{R}^d)$ denotes the Borel σ-algebra on $\mathbb{R}^d$. For each operator $o \in \mathcal{O}$ and base state $h \in H$, let

$$A_o(h) \in \mathcal{H}$$

denote the measurable set of admissible source states.

To allow source selection to be stochastic, we specify a source-selection rule by a Markov kernel

$$\nu_o : H \times \mathcal{H} \longrightarrow [0, 1]$$

such that

$$\nu_o(A_o(h) \mid h) = 1$$

for every $h \in H$. Thus $\nu_o(\cdot \mid h)$ is a probability distribution over admissible sources when the current base state is h.

If

$$S \sim \nu_o(\cdot \mid h),$$

then applying operator o produces the intervened neural state

$$H' = \Phi_o(h, S).$$

We therefore define the induced neural transition kernel by

$$\widehat{K}_o(h, B) := \mathbb{P}_{S \sim \nu_o(\cdot|h)}\big(\Phi_o(h, S) \in B\big) = \int_H \mathbf{1}_B(\Phi_o(h, s))\ \nu_o(ds \mid h), \qquad B \in \mathcal{H}.$$

Thus $\widehat{K}_o(h, \cdot)$ is the distribution of the next neural state obtained by applying operator o at the current state h.

We assume that

$$\Phi_o : H \times H \longrightarrow H$$

is measurable. Together with the Markov-kernel property of ν_o, this implies that

$$\widehat{K}_o : H \times \mathcal{H} \longrightarrow [0, 1]$$

is itself a Markov transition kernel on H; the verification is given in Appendix B.1.2.

In the experiments, source selection is deterministic. Writing a measurable selected-source map

$$s_o : H \longrightarrow H$$

for the selected-source map, with

$$s_o(h) \in A_o(h),$$

the corresponding source-selection kernel is

$$\nu_o(\cdot \mid h) = \delta_{s_o(h)},$$

where δ_x denotes the Dirac probability measure concentrated at x. Consequently,

$$\widehat{K}_o(h, \cdot) = \delta_{\Phi_o(h, s_o(h))}.$$

Hence the stochastic formulation above contains the deterministic intervention used in the experiments as a special case.

Let

$$S_0 \subseteq H$$

denote the set of neural states realized by the un-intervened evaluation distribution. We call S_0 the *observational support*.

The intervention closure is defined recursively by

$$\mathrm{Cl}_0(S_0) := S_0,$$

and, for $k \geq 1$,

$$\mathrm{Cl}_k(S_0) := \mathrm{Cl}_{k-1}(S_0) \cup \left\{ \Phi_o(h, s) : \begin{array}{l} h \in \mathrm{Cl}_{k-1}(S_0), \\ o \in \mathcal{O} \text{ is valid at } \widehat{\pi}(h), \\ s \in A_o(h) \end{array} \right\}.$$

For the deterministic benchmark protocol, admissible sources are un-intervened matched examples, so

$$A_o(h) \subseteq S_0.$$

Under this specialization, the recursive closure above agrees with the benchmark closure defined in §3.2.

Thus $\mathrm{Cl}_k(S_0)$ contains every neural state that can be reached from an observational state by at most k valid interventions using admissible sources.

In general,

$$S_0 = \mathrm{Cl}_0(S_0) \subseteq \mathrm{Cl}_1(S_0) \subseteq \cdots \subseteq \mathrm{Cl}_K(S_0),$$

and these inclusions may be strict. In particular, an intervention can produce a neural state that is never realized by the model under the original observational distribution. This is why correctness on S_0 alone does not control the behavior of later operators in a composition.

A.2.2 Observable discrepancy and the one-step intervention error

Let

$$d_{\mathrm{obs}} : \overline{Z} \times \overline{Z} \longrightarrow [0, \infty)$$

be a metric measuring the discrepancy between two observable high-level states.

For the finite categorical state space used by the benchmark, we use the discrete metric

$$d_{\mathrm{obs}}(z, z') := \mathbf{1}\{z \neq z'\} = \begin{cases} 0, & z = z', \\ 1, & z \neq z'. \end{cases}$$

We state the composition theorem for a general metric d_{obs}, while the discrete choice above gives the direct connection to observable disagreement and CIIA.

We continue to write

$$o : \mathrm{dom}(o) \subseteq \overline{Z} \longrightarrow \overline{Z}$$

for the operator induced on the observable quotient.

For a base neural state $h \in H$, the transition kernel $\widehat{K}_o(h, \cdot)$ gives the distribution of the next neural state after applying operator o. Pushing this distribution through the observable decoder $\widehat{\pi} : H \to \overline{Z}$ gives the corresponding distribution over observable high-level states, denoted by $\widehat{\pi}_{\#}\widehat{K}_o(h, \cdot)$.

For every measurable set $C \subseteq \overline{Z}$,

$$\big(\widehat{\pi}_{\#}\widehat{K}_o(h, \cdot)\big)(C) = \widehat{K}_o\big(h, \widehat{\pi}^{-1}(C)\big) = \int_H \mathbf{1}_C(\widehat{\pi}(\Phi_o(h, s)))\ \nu_o(ds \mid h).$$

Thus $\widehat{\pi}_{\#}\widehat{K}_o(h, \cdot)$ is the probability distribution of the decoded observable state after one intervention.

Because the high-level operator o is deterministic, the intended high-level next state is the single point

$$o(\widehat{\pi}(h)) \in \overline{Z}.$$

Its probability law is therefore the Dirac measure $\delta_{o(\widehat{\pi}(h))}$.

Let $\mathcal{P}(\overline{Z})$ denote the probability measures on $\overline{Z}$. For $\mu, \nu \in \mathcal{P}(\overline{Z})$, let $\Gamma(\mu, \nu)$ be the set of their couplings and define

$$W_1(\mu, \nu) := \inf_{\gamma \in \Gamma(\mu, \nu)} \int_{\overline{Z} \times \overline{Z}} d_{\mathrm{obs}}(z, z')\, d\gamma(z, z').$$

For a base neural state $h \in H$ and an operator o valid at $\widehat{\pi}(h)$, the decoded one-step law is $\widehat{\pi}_{\#}\widehat{K}_o(h,\cdot)$, while the deterministic high-level target is $o(\widehat{\pi}(h))$, with law $\delta_{o(\widehat{\pi}(h))}$. We therefore measure the one-step intervention error by

$$W_1\Big(\widehat{\pi}_{\#}\widehat{K}_o(h,\cdot), \delta_{o(\widehat{\pi}(h))}\Big).$$

Since the target law is a point mass, this quantity is exactly

$$W_1\Big(\widehat{\pi}_{\#}\widehat{K}_o(h,\cdot), \delta_{o(\widehat{\pi}(h))}\Big) = \int_H d_{\mathrm{obs}}(\widehat{\pi}(\Phi_o(h,s)), o(\widehat{\pi}(h)))\; d\nu_o(s \mid h).$$

Thus it is the expected observable discrepancy produced by one stochastic intervention. Under the discrete metric $d_{\mathrm{obs}}(z,z') = \mathbf{1}\{z \neq z'\}$, it reduces to

$$\mathbb{P}_{S\sim\nu_o(\cdot|h)}\left(\widehat{\pi}(\Phi_o(h,S)) \neq o(\widehat{\pi}(h))\right).$$

The derivation of these identities is given in Appendix B.1.5.

A.2.3 Admissible intervention processes

The process in Theorem 1 has neural states H_i, ideal high-level states Z_i, and selected operators O_i. Its history is specified by the following filtration.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the underlying probability space. For an intervention process of length $k \leq K$, define the filtration $(\mathcal{F}_i)_{i=0}^k$ by

$$\mathcal{F}_0 := \sigma(H_0, Z_0),$$

and, for $i \geq 1$,

$$\mathcal{F}_i := \sigma(H_0, Z_0, O_1, H_1, Z_1, \ldots, O_i, H_i, Z_i).$$

Thus $\mathcal{F}_i$ contains all information revealed up to and including the ith intervention.

For each $i = 1, \ldots, k$, the theorem requires O_i to be $\mathcal{F}_{i-1}$-measurable, so the operator choice can depend on the intervention history. The conditional transition requirement is $\mathcal{L}(H_i \mid \mathcal{F}_{i-1}) = \widehat{K}_{O_i}(H_{i-1},\cdot)$ almost surely. Equivalently, for every $B \in \mathcal{H}$,

$$\mathbb{P}(H_i \in B \mid \mathcal{F}_{i-1}) = \widehat{K}_{O_i}(H_{i-1}, B) \qquad \text{a.s.}$$

B Proofs and counterexamples

Appendix B.1.1 gives the composition proof, its deterministic specialization, and the reference-state error term. The supporting derivations establish the transition kernel (Appendix B.1.2), verify the finite witness and its enumeration (Appendices B.1.3 and B.1.4), and derive the one-step Wasserstein identity (Appendix B.1.5). Proposition 2 records the classical Occam bound. Appendix B.2 examines the roles of validity and closure-level control, the Lipschitz dependence, and tightness.

B.1 Proofs

B.1.1 Proof of Theorem 1

Proof of Theorem 1. Set

$$e_i := \mathbb{E}[d_{\mathrm{obs}}(\widehat{\pi}(H_i), Z_i)].$$

Since $Z_0 = \widehat{\pi}(H_0)$ almost surely, $e_0 = 0$.

By induction on i, the support condition $\nu_o(A_o(h) \mid h) = 1$ and admissibility of each realized intervention imply

$$H_i \in \mathrm{Cl}_i(S_0)$$

almost surely for every $0 \leq i \leq k$.

Fix $1 \le i \le k$. Since O_i is $\mathcal{F}_{i-1}$-measurable, the quantities H_{i-1}, Z_{i-1}, and O_i are determined by the intervention history up to time $i-1$. Moreover,

$$H_{i-1} \in \mathrm{Cl}_{i-1}(S_0) \subseteq \mathrm{Cl}_{K-1}(S_0)$$

almost surely, so assumption (A2) applies at the current neural state.

Conditioning on $\mathcal{F}_{i-1}$, the only new randomness in the one-step discrepancy comes from H_i. Therefore

$$\begin{aligned}
&\mathbb{E}[d_{\mathrm{obs}}(\widehat{\pi}(H_i), Z_i) \,|\, \mathcal{F}_{i-1}] \\
&= \mathbb{E}[d_{\mathrm{obs}}(\widehat{\pi}(H_i), O_i(Z_{i-1})) \,|\, \mathcal{F}_{i-1}] \\
&= \int_H d_{\mathrm{obs}}(\widehat{\pi}(h'), O_i(Z_{i-1}))\ \widehat{K}_{O_i}(H_{i-1}, dh').
\end{aligned}$$

The last equality follows from the conditional-law identity

$$\mathbb{E}[f(H_i) \,|\, \mathcal{F}_{i-1}] = \int_H f(h')\, \widehat{K}_{O_i}(H_{i-1}, dh'),$$

applied, conditional on $\mathcal{F}_{i-1}$, to

$$f(h') = d_{\mathrm{obs}}(\widehat{\pi}(h'), O_i(Z_{i-1}))\,.$$

By the triangle inequality,

$$\begin{aligned}
&d_{\mathrm{obs}}(\widehat{\pi}(h'), O_i(Z_{i-1})) \\
&\le d_{\mathrm{obs}}(\widehat{\pi}(h'), O_i(\widehat{\pi}(H_{i-1}))) \\
&\quad + d_{\mathrm{obs}}(O_i(\widehat{\pi}(H_{i-1})), O_i(Z_{i-1}))\,.
\end{aligned}$$

Assumption (A2) bounds the integral of the first term by ε, while (A1) bounds the second term by

$$L\, d_{\mathrm{obs}}(\widehat{\pi}(H_{i-1}), Z_{i-1})\,.$$

Hence

$$\mathbb{E}[d_{\mathrm{obs}}(\widehat{\pi}(H_i), Z_i) \,|\, \mathcal{F}_{i-1}] \le \varepsilon + L\, d_{\mathrm{obs}}(\widehat{\pi}(H_{i-1}), Z_{i-1})\,.$$

Taking expectations yields

$$e_i \le \varepsilon + L e_{i-1}.$$

Iterating gives

$$e_k \le L^k e_0 + \varepsilon \sum_{j=0}^{k-1} L^j,$$

and $e_0 = 0$ because $Z_0 = \widehat{\pi}(H_0)$ almost surely, which proves (6).

Finally, define the joint endpoint law

$$\gamma_k := \mathcal{L}(\widehat{\pi}(H_k), Z_k)\,.$$

Since its marginals are $\mathcal{L}(\widehat{\pi}(H_k))$ and $\mathcal{L}(Z_k)$, we have

$$\gamma_k \in \Gamma(\mathcal{L}(\widehat{\pi}(H_k)), \mathcal{L}(Z_k))\,.$$

Therefore, by the definition of the Wasserstein distance,

$$\begin{aligned}
W_1(\mathcal{L}(\widehat{\pi}(H_k)), \mathcal{L}(Z_k)) &\le \int_{\overline{Z}\times\overline{Z}} d_{\mathrm{obs}}(z, z')\, d\gamma_k(z, z') \\
&= \mathbb{E}[d_{\mathrm{obs}}(\widehat{\pi}(H_k), Z_k)]\,.
\end{aligned}$$

Combining this with (6) gives

$$W_1(\mathcal{L}(\widehat{\pi}(H_k)), \mathcal{L}(Z_k)) \le \varepsilon \sum_{j=0}^{k-1} L^j,$$

which proves (7).

□

Corollary 1 (Deterministic intervention protocol)**.** *Suppose the operator word is fixed and each operator o uses a deterministic source-selection map $s_o : H \to H$. Then*

$$\widehat{K}_o(h, \cdot) = \delta_{\Phi_o(h, s_o(h))}.$$

In this case, the one-step condition (5) *reduces to*

$$d_{\mathrm{obs}}(\widehat{\pi}(\Phi_o(h, s_o(h))), o(\widehat{\pi}(h))) \le \varepsilon$$

for every admissible $h \in \mathrm{Cl}_{K-1}(S_0)$.

Hence Theorem 1 specializes to the deterministic composition bound

$$d_{\mathrm{obs}}(\widehat{\pi}(H_k), Z_k) \le \varepsilon \sum_{j=0}^{k-1} L^j.$$

In particular, when $L = 1$,

$$d_{\mathrm{obs}}(\widehat{\pi}(H_k), Z_k) \le k\varepsilon.$$

Thus the stochastic formulation extends, rather than alters, the deterministic intervention protocol used in the experiments.

Remark 2 (Scoring against a reference state). The experiments score the decoded endpoint against $w(z_b)$ for the generator's state z_b of the base example, not against $w(\widehat{\pi}(H_0))$. Running the recursion $e_i \le \varepsilon + Le_{i-1}$ from $Z_0 = z_b$ instead, with $e_0 := \mathbb{E}[d_{\mathrm{obs}}(\widehat{\pi}(H_0), z_b)]$ the observational decoding error at the base, gives

$$\mathbb{E}[d_{\mathrm{obs}}(\widehat{\pi}(H_k), Z_k)] \le L^k e_0 + \varepsilon \sum_{j=0}^{k-1} L^j$$

under the same validity conditions, which must now hold along both trajectories. The identification of the left side with $1 - \mathrm{CIIA}_k$ in §4 is therefore exact on bases the model decodes correctly and acquires the term $L^k e_0$ otherwise; §6 reports e_0 on the replication organisms' closure and unary bases.

B.1.2 Transition kernel

Lemma 1 (The induced intervention law is a Markov transition kernel)**.** *Let $(H, \mathcal{H})$ be the neural state space. Let*

$$\nu_o : H \times \mathcal{H} \longrightarrow [0, 1]$$

be a Markov kernel specifying the source-selection rule for operator o, and assume that

$$\Phi_o : H \times H \longrightarrow H$$

is measurable. Define

$$\widehat{K}_o(h, B) := \int_H \mathbf{1}_B(\Phi_o(h, s))\ \nu_o(ds \mid h), \qquad h \in H, \quad B \in \mathcal{H}.$$

Then

$$\widehat{K}_o : H \times \mathcal{H} \longrightarrow [0, 1]$$

is a Markov transition kernel on H.

Proof. We verify the two defining properties of a Markov transition kernel.

First, fix $h \in H$ and define

$$T_h : H \longrightarrow H, \qquad T_h(s) := \Phi_o(h, s).$$

Since Φ_o is measurable, T_h is measurable. For every $B \in \mathcal{H}$,

$$\widehat{K}_o(h, B) = \nu_o\big(T_h^{-1}(B) \mid h\big)\,.$$

Equivalently,

$$\widehat{K}_o(h, \cdot) = (T_h)_{\#}\nu_o(\cdot \mid h),$$

where $(T_h)_\#$ denotes pushforward of measures under T_h.

For fixed h, $\nu_o(\cdot \mid h)$ is a probability measure on $(H, \mathcal{H})$, and the pushforward of a probability measure under a measurable map is again a probability measure. Hence

$$B \longmapsto \widehat{K}_o(h, B)$$

is a probability measure on $(H, \mathcal{H})$.

Second, fix $B \in \mathcal{H}$. Since Φ_o is measurable, the function

$$(h, s) \longmapsto \mathbf{1}_B(\Phi_o(h, s))$$

is measurable on $H \times H$. Because ν_o is a Markov kernel, the standard measurability theorem for integration against a kernel implies that

$$h \longmapsto \int_H \mathbf{1}_B(\Phi_o(h, s))\ \nu_o(ds \mid h)$$

is measurable. Therefore

$$h \longmapsto \widehat{K}_o(h, B)$$

is measurable.

Thus, for every fixed h, $\widehat{K}_o(h, \cdot)$ is a probability measure, and for every fixed $B \in \mathcal{H}$, $\widehat{K}_o(\cdot, B)$ is measurable. Therefore $\widehat{K}_o$ is a Markov transition kernel on H. □

B.1.3 THE FINITE WITNESS

A concrete witness is obtained by taking

$$Z = \{0, 1\}^2, \qquad H = \{0, 1\}^3,$$

with

$$\widehat{\pi}(v_1, v_2, v_3) = (v_1, v_2), \qquad S_0 = \{(v_1, v_2, v_3) \in H : v_3 = v_1\}.$$

Define

$$a(x_1, x_2) = (1, x_2), \qquad b(x_1, x_2) = (x_1, x_1),$$

and

$$\Phi_a(v_1, v_2, v_3) = (1, v_2, v_3), \qquad \Phi_b(v_1, v_2, v_3) = (v_1, v_3, v_3).$$

Both unary implementations are exact on S_0, while for

$$h_0 = (0, 0, 0) \in S_0$$

the composition satisfies

$$\widehat{\pi}\big((\Phi_b \circ \Phi_a)(h_0)\big) \neq (b \circ a)\big(\widehat{\pi}(h_0)\big),$$

with

$$\Phi_a(h_0) \in \mathrm{Cl}_1(S_0) \setminus S_0.$$

The full verification is given in Appendix B.1.4.

B.1.4 VERIFICATION OF THE UNARY-INSUFFICIENCY COUNTEREXAMPLE

Take

$$Z = \{0, 1\}^2, \qquad H = \{0, 1\}^3,$$

with

$$\widehat{\pi}(v_1, v_2, v_3) = (v_1, v_2)$$

and

$$S_0 = \{(v_1, v_2, v_3) \in H : v_3 = v_1\}.$$

Define

$$a(x_1, x_2) = (1, x_2), \qquad b(x_1, x_2) = (x_1, x_1),$$

and

$$\Phi_a(v_1, v_2, v_3) = (1, v_2, v_3), \qquad \Phi_b(v_1, v_2, v_3) = (v_1, v_3, v_3).$$

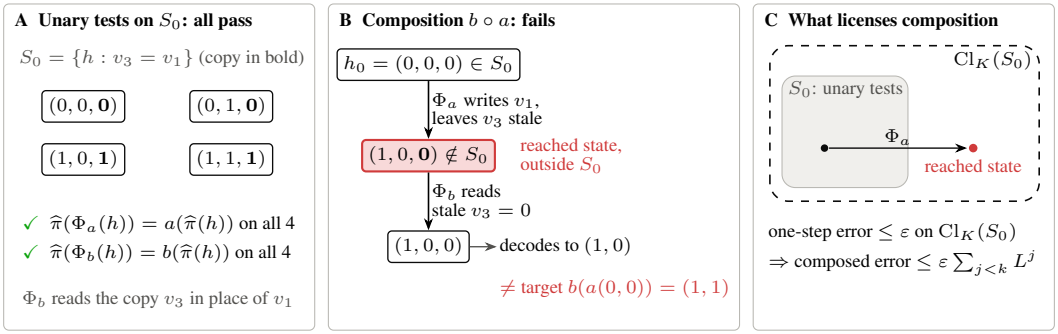


Figure 2: **Unary correctness and a failed composition path.** Panel A shows eight successful unary checks across the four observational states, where the redundant copy agrees with the first coordinate. Panel B follows a path in which the first intervention changes that coordinate while leaving the copy unchanged, and the second reads the stale copy to produce the wrong endpoint. Panel C sketches closure-level error control. Theorem 1 and Appendix B.1.1 give the full assumptions and proof.

For every $h = (v_1, v_2, v_3) \in S_0$,

$$\widehat{\pi}(\Phi_a(h)) = (1, v_2) = a(\widehat{\pi}(h)).$$

Moreover, since $v_3 = v_1$ on S_0,

$$\widehat{\pi}(\Phi_b(h)) = (v_1, v_3) = (v_1, v_1) = b(\widehat{\pi}(h)).$$

Thus both unary implementations are exact on the observational support.

Now take

$$h_0 = (0, 0, 0) \in S_0.$$

At the high level,

$$(b \circ a)(\widehat{\pi}(h_0)) = b(a(0, 0)) = b(1, 0) = (1, 1).$$

At the low level,

$$\Phi_a(h_0) = (1, 0, 0),$$

and hence

$$(\Phi_b \circ \Phi_a)(h_0) = \Phi_b(1, 0, 0) = (1, 0, 0).$$

Therefore

$$\widehat{\pi}\big((\Phi_b \circ \Phi_a)(h_0)\big) = (1, 0) \neq (1, 1) = (b \circ a)(\widehat{\pi}(h_0)).$$

Finally,

$$\Phi_a(h_0) = (1, 0, 0) \in \mathrm{Cl}_1(S_0),$$

but

$$(1, 0, 0) \notin S_0$$

because its first and third coordinates are unequal. Hence

$$\Phi_a(h_0) \in \mathrm{Cl}_1(S_0) \setminus S_0.$$

Thus the compositional failure occurs precisely because the second implementation is applied at an intervention-reached state outside the observational support.

Machine check. `scripts/verify_counterexample.py` performs the same enumeration mechanically: all 8 unary checks on $S_0 = \{(0, 0, 0), (0, 1, 0), (1, 0, 1), (1, 1, 1)\}$ pass, and of the 4 words of length 2 on the 4 support states exactly two fail, both for $b \circ a$ (base states $(0, 0, 0)$ and $(0, 1, 0)$), each at an intermediate state in $\mathrm{Cl}_1(S_0) \setminus S_0$.

Necessity of each ingredient. If $S_0 = H$ (no off-support states), unary exactness forces Φ_b correct everywhere and composition succeeds; if v_3 is removed (no redundant correlate), no implementation of b that is exact on S_0 can fail under composition.

On minimality. Exhaustive enumeration considers all systems with two total operators, a surjective $\widehat{\pi}$, and a nonempty S_0 (`verify_review_claims.py`). It identifies the following minima under two sets of constraints.

- without further constraints the minimum is $|Z| = |H| = 2$, with S_0 a single state—but this witness is *degenerate*: it fails only because b was never tested at the other high-level state, which is not the phenomenon of interest;
- under the non-degeneracy constraint $\widehat{\pi}(S_0) = Z$ (every high-level state *is* observationally attested, so no failure can be attributed to an untested high-level state) the minimum is $|Z| = 2$, $|H| = 3$, and no system with $|H| = 2$ admits a counterexample.

The $|Z| = 4$, $|H| = 8$ construction illustrates a stale agreement correlate through v_3.

B.1.5 One-step Wasserstein discrepancy

We justify the one-step Wasserstein identity used in the main text. Fix a base neural state $h \in H$ and an operator o valid at $\widehat{\pi}(h)$. For convenience, write

$$\nu_o^h := \nu_o(\cdot \mid h), \qquad \kappa_o^h := \widehat{K}_o(h, \cdot),$$

and define

$$\mu_o^h := \widehat{\pi}_{\#}\kappa_o^h, \qquad z_o(h) := o(\widehat{\pi}(h)).$$

Thus ν_o^h is the source-selection law, κ_o^h is the induced next-state law on H, μ_o^h is the corresponding decoded law on $\overline{Z}$, and $z_o(h)$ is the intended deterministic high-level target.

Lemma 2 (One-step Wasserstein identity)**.** *For every such h and o,*

$$W_1\Big(\widehat{\pi}_{\#}\widehat{K}_o(h, \cdot), \delta_{o(\widehat{\pi}(h))}\Big) = \int_H d_{\mathrm{obs}}(\widehat{\pi}(\Phi_o(h, s)), o(\widehat{\pi}(h)))\; d\nu_o^h(s).$$

If d_{obs} is the discrete metric

$$d_{\mathrm{obs}}(z, z') = \mathbf{1}\{z \neq z'\},$$

then

$$W_1\Big(\widehat{\pi}_{\#}\widehat{K}_o(h, \cdot), \delta_{o(\widehat{\pi}(h))}\Big) = \mathbb{P}_{S \sim \nu_o^h}\left(\widehat{\pi}(\Phi_o(h, S)) \neq o(\widehat{\pi}(h))\right).$$

Proof. By definition,

$$W_1\big(\mu_o^h, \delta_{z_o(h)}\big) = \inf_{\gamma \in \Gamma(\mu_o^h, \delta_{z_o(h)})} \int_{\overline{Z} \times \overline{Z}} d_{\mathrm{obs}}(z, z')\, d\gamma(z, z').$$

For any

$$\gamma \in \Gamma(\mu_o^h, \delta_{z_o(h)}),$$

the second marginal of γ is the point mass $\delta_{z_o(h)}$. Hence the second coordinate equals $z_o(h)$ γ-almost surely. Therefore

$$\int_{\overline{Z} \times \overline{Z}} d_{\mathrm{obs}}(z, z')\, d\gamma(z, z') = \int_{\overline{Z} \times \overline{Z}} d_{\mathrm{obs}}(z, z_o(h))\; d\gamma(z, z').$$

The integrand on the right depends only on the first coordinate, whose marginal under γ is μ_o^h. Hence

$$\int_{\overline{Z} \times \overline{Z}} d_{\mathrm{obs}}(z, z_o(h))\; d\gamma(z, z') = \int_{\overline{Z}} d_{\mathrm{obs}}(z, z_o(h))\; d\mu_o^h(z).$$

This value is independent of the choice of coupling γ, so taking the infimum gives

$$W_1\big(\mu_o^h, \delta_{z_o(h)}\big) = \int_{\overline{Z}} d_{\mathrm{obs}}(z, z_o(h))\; d\mu_o^h(z).$$

Since

$$\mu_o^h = \widehat{\pi}_{\#}\kappa_o^h,$$

the defining integration identity for a pushforward measure gives, for every nonnegative measurable or integrable function $f : \overline{Z} \to \mathbb{R}$,

$$\int_{\overline{Z}} f(z)\, d\mu_o^h(z) = \int_H f(\widehat{\pi}(h'))\, d\kappa_o^h(h').$$

For completeness, the identity first holds for indicator functions by the definition of pushforward measure. By linearity, it therefore holds for every nonnegative simple function. If $f_n \uparrow f$ is an increasing sequence of nonnegative simple functions converging pointwise to a nonnegative measurable function f, then the monotone convergence theorem gives

$$\int_{\overline{Z}} f \, d\mu_o^h = \lim_{n\to\infty} \int_{\overline{Z}} f_n \, d\mu_o^h = \lim_{n\to\infty} \int_H f_n(\widehat{\pi}(h')) \, d\kappa_o^h(h') = \int_H f(\widehat{\pi}(h')) \, d\kappa_o^h(h').$$

Hence the pushforward integration identity holds for every nonnegative measurable f. If f is integrable, applying the same identity to its positive and negative parts,

$$f = f^+ - f^-,$$

shows that it also holds for every integrable measurable function f.

Taking

$$f(z) = d_{\mathrm{obs}}(z, z_o(h))$$

yields

$$W_1\big(\mu_o^h, \delta_{z_o(h)}\big) = \int_H d_{\mathrm{obs}}(\widehat{\pi}(h'), z_o(h)) \; d\kappa_o^h(h').$$

Using

$$z_o(h) = o(\widehat{\pi}(h)),$$

this becomes

$$W_1\Big(\widehat{\pi}_{\#}\widehat{K}_o(h,\cdot), \delta_{o(\widehat{\pi}(h))}\Big) = \int_H d_{\mathrm{obs}}(\widehat{\pi}(h'), o(\widehat{\pi}(h))) \; d\kappa_o^h(h').$$

For fixed h, define

$$T_{o,h} : H \longrightarrow H, \qquad T_{o,h}(s) := \Phi_o(h, s).$$

By the definition of the induced intervention kernel,

$$\kappa_o^h = (T_{o,h})_{\#}\nu_o^h.$$

Applying the pushforward integration identity once more gives

$$\int_H g(h') \, d\kappa_o^h(h') = \int_H g(\Phi_o(h,s)) \, d\nu_o^h(s)$$

for every nonnegative measurable or integrable $g : H \to \mathbb{R}$. Taking

$$g(h') = d_{\mathrm{obs}}(\widehat{\pi}(h'), o(\widehat{\pi}(h)))$$

therefore gives

$$W_1\Big(\widehat{\pi}_{\#}\widehat{K}_o(h,\cdot), \delta_{o(\widehat{\pi}(h))}\Big) = \int_H d_{\mathrm{obs}}(\widehat{\pi}(\Phi_o(h,s)), o(\widehat{\pi}(h))) \; d\nu_o^h(s).$$

Finally, under the discrete metric,

$$d_{\mathrm{obs}}(\widehat{\pi}(\Phi_o(h,s)), o(\widehat{\pi}(h))) = \mathbf{1}\left\{\widehat{\pi}(\Phi_o(h,s)) \neq o(\widehat{\pi}(h))\right\}.$$

Hence

$$\begin{aligned} &W_1\Big(\widehat{\pi}_{\#}\widehat{K}_o(h,\cdot), \delta_{o(\widehat{\pi}(h))}\Big) \\ &= \int_H \mathbf{1}\left\{\widehat{\pi}(\Phi_o(h,s)) \neq o(\widehat{\pi}(h))\right\} \, d\nu_o^h(s) \\ &= \mathbb{P}_{S\sim\nu_o^h}\left(\widehat{\pi}(\Phi_o(h,S)) \neq o(\widehat{\pi}(h))\right). \end{aligned}$$

Thus, for the discrete metric, the one-step Wasserstein discrepancy is exactly the probability that the decoded neural intervention disagrees with its intended high-level target. □

B.1.6 EXPLANATION SELECTION: THE STANDARD OCCAM BOUND

Proposition 2 (Occam bound; classical, included for completeness only)**.** *Let Π be a countable class of explanations with prefix code lengths $\mathrm{DL}(\pi)$ (in bits). Draw n base examples i.i.d. from the base marginal of $\mathcal{D}_k$; for each base, let its loss in $[0,1]$ be the conditional mean over the declared within-base word/source law. Let $\hat{\mathcal{L}}_n(\pi)$ be the empirical mean of those base-level losses and $\mathcal{L}(\pi)$ its population mean. With probability at least $1-\delta$, simultaneously for all $\pi \in \Pi$:*

$$\mathcal{L}(\pi) \;\le\; \hat{\mathcal{L}}_n(\pi) + \sqrt{\frac{\mathrm{DL}(\pi)\ln 2 + \ln(1/\delta)}{2n}}.$$

Proof. Hoeffding's inequality for each π at level $\delta\, 2^{-\mathrm{DL}(\pi)}$, then a union bound using Kraft's inequality $\sum_\pi 2^{-\mathrm{DL}(\pi)} \le 1$. □ □

This classical bound controls the population mean of the declared base-level loss by its empirical mean and a penalty determined by code length, sample size, and confidence level. The guarantee holds uniformly over the prefix-coded explanation class. Each loss averages over the declared within-base word/source law, with base examples sampled i.i.d. The experimental comparisons use the same base-example unit for resampling (Appendix E.1).

B.2 COUNTEREXAMPLES FOR REMOVED ASSUMPTIONS

Each counterexample below is executed numerically by the script `verify_metric_definitions.py`; we record the constructions here. They concern the deterministic form of the bound, Corollary 1, whose one-step condition is

$$d_{\mathrm{obs}}\big(\widehat{\pi}(\Phi_o(h, s_o(h))),\, o(\widehat{\pi}(h))\big) \le \varepsilon \qquad \text{at every admissible } h \in \mathrm{Cl}_{K-1}(S_0).$$

B.2.1 VALIDITY AT THE DECODED STATE REMOVED: THE BOUND FAILS

Let $Z = H = \mathbb{R}$ with the usual metric and $\widehat{\pi} = \mathrm{id}$. Define two partial generators with single-point domains,

$$o_1 : \{0\} \to \mathbb{R},\; o_1(0) = 1, \qquad o_2 : \{1\} \to \mathbb{R},\; o_2(1) = 0,$$

so the word $o_2 \circ o_1$ is valid at $z = 0$, and both generators are L-Lipschitz for every L (any map on a single point is). Let $\Phi_1(h) = h + 1 + \delta$ with $\delta = 0.05$ and $\Phi_2(1) = 0$, $\Phi_2(h) = h + 10^6$ for $h \neq 1$. The one-step condition holds with $\varepsilon = 0.1$ at every state where the corresponding generator is *valid*: $|\Phi_1(0) - o_1(0)| = \delta$ and $|\Phi_2(1) - o_2(1)| = 0$. But $\widehat{\pi}(\Phi_1(0)) = 1.05 \notin \mathrm{dom}(o_2)$: the trajectory $H_0 = 0$, $H_1 = \Phi_1(0)$, $O_2 = o_2$ is not an admissible intervention process in the sense of Theorem 1, the one-step condition is vacuous at H_1, and

$$d_{\mathrm{obs}}\big(\widehat{\pi}(\Phi_2(\Phi_1(0))),\; o_2(o_1(0))\big) = 1{,}000{,}001.05 \;\gg\; \varepsilon(1+L) = 0.2.$$

Hence the admissibility requirement—validity of each operator at the decoded intervened state $\widehat{\pi}(H_{i-1})$, not only at the ideal state Z_{i-1}—cannot be dropped: without it the conclusion fails by an arbitrary factor. Verified numerically in `verify_review_claims.py`, which also checks the bound on $400/400$ random systems whose accumulated error stays inside a validity margin, so that every step remains valid at the decoded state.

B.2.2 ONE-STEP CONDITION ON S_0 ONLY (CLOSURE CLAUSE REMOVED)

Take the finite system of Proposition 1 and equip Z with the discrete metric $d_{\mathrm{obs}}(z, z') = \mathbf{1}\{z \neq z'\}$. Every generator is 1-Lipschitz (any map is, under the discrete metric). The unary discrepancy on S_0 is exactly 0, so if the one-step condition restricted to S_0 sufficed, the bound (6) would force composed error $\le 2 \cdot 0 = 0$. The observed composed error is 1 (Proposition 1). The condition indeed fails on the closure: at $(1,0,0) \in \mathrm{Cl}_1(S_0)$ the unary discrepancy of Φ_b is 1. Hence the closure clause in (A2) cannot be weakened to S_0.

B.2.3 DEPENDENCE ON THE LIPSCHITZ CONSTANT

Take $Z = H = \mathbb{R}$, $\widehat{\pi} = \mathrm{id}$, one generator $o(z) = 2z$ ($L = 2$), and $\Phi(h) = 2h + \varepsilon$ with $\varepsilon = 0.1$. The one-step condition holds everywhere with tolerance ε. After $k = 4$ compositions the error is $\varepsilon(2^4 - 1) = 1.5$, exceeding the $L{=}1$ bound $k\varepsilon = 0.4$, while matching the correct bound $\varepsilon \sum_{j=0}^{3} 2^j$ exactly. Hence the L-dependence of (6) is real and the bound cannot be stated uniformly in L.

B.2.4 TIGHTNESS OF THE BOUND

With $o(z) = Lz$ and $\Phi(h) = Lh + \varepsilon$, induction gives $\widehat{\pi}(\Phi^k(h)) - o^k(\widehat{\pi}(h)) = \varepsilon \sum_{j=0}^{k-1} L^j$ for every h, so the bound of Corollary 1 holds with equality for all k and all $L > 0$: no smaller coefficient sequence is possible without further assumptions.

B.2.5 ROLES OF THE ASSUMPTIONS

The validity construction gives arbitrarily large composed error when the neural trajectory reaches a decoded state outside the next high-level operator's domain, even though the ideal sequence remains valid. The finite witness shows that exact one-step control on the observational support can coexist with composed error once the trajectory leaves that support. The linear constructions show how the Lipschitz constant determines error propagation and attain the composition bound with equality. Section 6 reports the empirical behavior of fitted interventions.

C EFFECTIVE DEPTH AND BENCHMARK CONSTRUCTION

Effective depth connects the length of an operator sequence to the shortest valid sequence reaching the same observable endpoint from the same base. Appendix C.1 gives the definition and score decomposition, and Appendix C.2 describes how OpClosure-v2 filters candidate words before sampling. The grammar and its realization rules follow in Appendix C.3. The OpClosure development records and depth analysis appear in Appendices C.3.5 and C.3.6.

C.1 EFFECTIVE COMPOSITION DEPTH

The composition bound in Theorem 1 controls error along a sequence of a given length. Interpreting the depth evaluations in Section 5.1 also requires the shortest valid word reaching the same observable endpoint from the same base. Definition 2 makes this effective depth precise, and Proposition 3 decomposes accuracy when a pool contains both genuine and endpoint-reducible rows.

Let

$$q : Z \to \overline{Z}, \qquad q(z) := [z]_{\mathcal{C}},$$

denote the canonical quotient map. Fix a base state $z \in Z$. For every valid word $w \in \mathcal{W}_{\leq K}(z)$, define its observable endpoint by

$$\mathrm{Ev}_z(w) := q(w(z)) \in \overline{Z}.$$

The set of observable endpoints reachable from z within depth K is

$$\mathcal{R}_K(z) := \{\mathrm{Ev}_z(w) : w \in \mathcal{W}_{\leq K}(z)\}.$$

Definition 2 (Effective composition depth)**.** For $z \in Z$ and $v \in \mathcal{R}_K(z)$, define the effective composition depth of v relative to z by

$$\mathrm{edepth}_z(v) := \min\{|w| : w \in \mathcal{W}_{\leq K}(z),\ \mathrm{Ev}_z(w) = v\}. \tag{8}$$

An observable target $v \in \mathcal{R}_K(z)$ is called a *genuine depth-k target* relative to z if

$$\mathrm{edepth}_z(v) = k.$$

A valid word $w \in \mathcal{W}_k(z)$ is called *endpoint-reducible at z* if

$$\mathrm{edepth}_z(\mathrm{Ev}_z(w)) < k.$$

Otherwise, w is called *endpoint-genuine at z*.

A length-k word is endpoint-genuine if its observable endpoint cannot be reached from the same base state by any shorter valid word. Thus formal word length can overstate effective depth when operator relations create redundant endpoints. For example, if

$$q(o \circ o(z)) = q(z),$$

then $o \circ o$ has formal length two but effective depth zero.

Effective depth is defined only at the level of observable endpoints: distinct intervention paths may have the same endpoint without being equivalent as neural trajectories, so the definition does not eliminate order or path dependence.

We next use this distinction to refine depth-k evaluation. Let (X_0, W) be sampled according to the declared depth-k evaluation law, with

$$X_0 \in Z, \qquad W \in \mathcal{W}_k(X_0)$$

almost surely. Define the event

$$G_k := \left\{ \mathrm{edepth}_{X_0}(\mathrm{Ev}_{X_0}(W)) = k \right\},$$

and let

$$\alpha_k := \mathbb{P}(G_k).$$

Thus α_k is the probability that a sampled formal depth-k intervention has an observable endpoint whose effective depth is genuinely k.

Let C_k denote the event that the model produces the correct observable outcome under the sampled depth-k intervention, and define

$$\mathrm{CIIA}_k := \mathbb{P}(C_k).$$

Whenever the corresponding conditioning events have positive probability, define

$$\mathrm{CIIA}_k^{\mathrm{gen}} := \mathbb{P}(C_k \mid G_k), \qquad \mathrm{CIIA}_k^{\mathrm{red}} := \mathbb{P}(C_k \mid G_k^c).$$

Proposition 3 (Effective-depth decomposition)**.** *Fix $1 \le k \le K$ and suppose*

$$0 < \alpha_k = \mathbb{P}(G_k) < 1.$$

Then

$$\mathrm{CIIA}_k = \alpha_k \, \mathrm{CIIA}_k^{\mathrm{gen}} + (1 - \alpha_k) \, \mathrm{CIIA}_k^{\mathrm{red}} . \tag{9}$$

Proof. Since G_k and G_k^c form a partition of the evaluation space,

$$\begin{aligned} \mathrm{CIIA}_k = \mathbb{P}(C_k) &= \mathbb{P}(C_k \mid G_k)\mathbb{P}(G_k) + \mathbb{P}(C_k \mid G_k^c)\mathbb{P}(G_k^c) \\ &= \alpha_k \, \mathrm{CIIA}_k^{\mathrm{gen}} + (1 - \alpha_k) \, \mathrm{CIIA}_k^{\mathrm{red}} . \end{aligned}$$

□

Equation (9) shows how pool composition affects depth accuracy. When endpoint-reducible rows have higher accuracy, their inclusion raises the pooled score relative to the genuine-depth score. Appendix C.3.6 reports the development-pool partition and its effect on the depth comparisons.

C.2 Endpoint-genuine pools in OpClosure-v2

OpClosure-v2 applies Definition 2 to candidate words before sampling, retaining words whose effective depth equals their length. The same predicate applies to the closure pool. Sampling therefore draws from candidates whose endpoints require the declared number of operations. Table 1 checks this property in the resulting pools. It reports 0 endpoint-reducible rows and 2163 genuine rows, including 105 genuine IID depth-3 rows, 67 genuine depth-4 rows, and 500 genuine closure rows.

C.3 The OpClosure synthetic grammar

The grammar specifies the state fields, surface realizations, and valid transformations that produce OpClosure's counterfactual targets. For intervention rows, the realization places grammatical cues before the site and an observable continuation after it. Appendix C.3.5 records the released development instance, its split rules, and validation results.

Table 1: Effective-depth census of the v2 pools. A row is endpoint-genuine when no strictly shorter valid word reaches the same observable endpoint (Definition 2), so a genuine depth-k row is a test of a depth-k composition and not of a shorter one. The v2 generator filters the candidate set through this predicate before sampling, so these pools are genuine by construction; the census is the check that they are.

Pool	rows	genuine	reducible	effective depths present
`test_unary`	500	500	0	d1: 500
`test_pairs_iid`	500	500	0	d2: 500
`test_pair`	264	264	0	d2: 264
`test_order`	110	110	0	d2: 110
`test_triples_iid`	105	105	0	d3: 105
`test_triple`	117	117	0	d3: 117
`test_depth`	67	67	0	d4: 67
`test_closure`	500	500	0	d2: 500
total reducible rows			0	

C.3.1 State space and lexicon

States are tuples $(\textit{subj}, \textit{num}, \textit{verb}, \textit{tense}, \textit{pol}, \textit{voice}, \textit{constr}, \textit{theme}, \textit{path}, \textit{gap})$ with $\textit{num} \in \{\text{SG}, \text{PL}\}$, $\textit{tense} \in \{\text{PRES}, \text{PAST}\}$, $\textit{pol} \in \{\text{POS}, \text{NEG}\}$, $\textit{voice} \in \{\text{ACT}, \text{PASS}\}$, $\textit{constr} \in \{\text{INTR}, \text{TRANS}, \text{CM}\}$, $\textit{gap} \in \{\text{NONE}, \text{WH_THEME}\}$. Themes are singular, so passive agreement is determined by the theme. Changing deep-agent number is surface-visible in actives and questions and leaves passive agreement unchanged. The lexicon (10 transitive verbs, 5 manner verbs, 10 agents, 10 themes, 4 caused-motion themes, 6 paths; all regular plurals; explicit form tables for irregular verb morphology) is partitioned into class A (training) and class B (lexical holdout); any example containing a class-B lexeme is excluded from training and validation pools entirely.

C.3.2 Realization: main clause, site, echo tail

Every declarative and passive has the form

$$\underbrace{\text{main clause}}_{\text{realizes every feature}} \quad \underbrace{,}_{\text{site}} \quad \underbrace{\text{[and the THEME is PROP] , tag question}}_{\text{reflects every feature}}$$

where the bracketed coordinated theme-property clause is present exactly when the state has a theme. For example, *the chefs did not eat the cake , and the cake is sweet , did they ?* The full echo is an engineered diagnostic continuation with English auxiliary, pronoun, agreement, and polarity morphology. It exposes the operator's effect after the intervention site. The interchange operator transports the difference between a matched source's and the base's activation at the site; if the operator's variable is not encoded in the prefix, the matched source and the base are token-identical up to the site, $h_s - h_b = 0$, and Φ_o is literally the identity map. In an earlier version of this benchmark, tense, polarity, and caused motion were realized *at* the finite verb rather than before it; their base/source prefixes were then token-identical, forcing the intervention to be the identity. The echo tail moves a reflex of each feature strictly after the site:

operator	cue (before the site)	reflex (after the site)
NUM	subject noun	tag pronoun and tag auxiliary
TENSE	finite verb / auxiliary	tag auxiliary
POL	*not* and do-support	*not* in the tag (polarity reversal)
PASS	clause structure	tag auxiliary family, *it* subject
BIND	theme noun	property adjective
CM	theme and path	presence of the coordinated theme-property clause

After redesign, the generator's self-test asserts, for every one of the six autoregressive candidate operators, that the realization differs both before and after the site, with the declared deep-agent-number case in passives masked. Development-suite validation excludes CM, leaving five active candidates. Polarity-reversing tag morphology supplies a linguistically interpretable reflex. The engineered echo serves as benchmark instrumentation.

Object wh-questions take no tag in English; they are generated as behavioural stimuli for the language model but carry no site and are never intervention bases, so WH remains frame-only under the autoregressive protocol (Appendix D.1).

The outcome signature is a congruence. Remark 1 requires the candidate-set equivalence to be preserved by the operators. An exhaustive check over the reachable-state fixed point finds no violations of this congruence condition.

C.3.3 SURFACE FORMS

A deterministic realization function produces: present-tense agreement (*eats/eat*), do-support under negation and in questions (*does/do/did*), passive auxiliaries with participles (*is/are/was/were* + *eaten*), object wh-questions with subject–aux inversion (*which cake did the chefs not eat?*), and caused-motion frames (*the critics laugh the actor off the stage*). The finite verb/aux is the *diagnostic token*: the unique position where number, tense, polarity, and voice jointly surface.

C.3.4 OPERATORS AND VALIDITY

The seven operators of §3.1, with validity: PASS requires TRANS, ACT, no gap; WH requires TRANS, ACT, no gap; CM requires INTR and a caused-motion-capable verb, and deterministically selects the alphabetically-first same-lexical-class theme and path; BIND requires an existing theme and rebinds to the next same-class theme (cycling within the lexical split class so BIND never crosses the lexical holdout boundary); the three feature toggles are total. Passivization of caused-motion frames and wh-extraction from passives are excluded by validity (a scope restriction, recorded in Appendix K.2). Invalid compositions are masked cells in every matrix, never zeros.

C.3.5 OPCLOSURE DEVELOPMENT RECORDS AND POOLS

The development release uses generator seed 17 and contains 108,362 intervention rows over 3,701 examples. Each evaluated row stores the latent state, operator sequence, matched sources, site, target state and surface realization, and pool assignment. Some auxiliary per-token bookkeeping fields are inconsistent with the realized strings in the development release; these fields are not used by fitting, evaluation, or any reported statistic.

Splits and pools. Example-level counts are 1,521 train, 493 validation, 615 test, and 1,072 lexical holdout; 2,880 examples carry an echo tail and can serve as intervention bases. The development suite used an earlier deletion-based degeneracy filter, so some sampled higher-depth rows are endpoint-reducible; Appendix C.3.6 quantifies this, while OpClosure-v2 filters by effective depth before sampling.

pool	role	rows
`analysis_only`	not evaluated (train-split rows of depth > 1)	31,276
`test_lex`	test, held-out lexicon	27,531
`excluded_degenerate`	not evaluated (a proper subword induces the same observable transformation, so the row cannot test the full word)	12,407
`train`	operator fitting (unary only)	6,802
`excluded`	not evaluated (validation-split held signatures)	6,435
`test_pairs_iid`	test, depth 2, seen signatures	3,539
`val_pairs`	validation, depth 2	3,339
`test_closure`	test, intervention closure	3,084
`test_surface`	test, surface perturbation	2,518
`test_unary`	test, depth 1	2,316
`val`	validation, unary	2,210
`test_depth`	test, depth 4	1,604
`test_pair`	test, depth 2, HELD-OUT signatures	1,548
`test_triples_iid`	test, depth 3, seen signatures	1,494
`test_triple`	test, depth 3, HELD-OUT signatures	1,017
`test_pair_sub`	test, depth 3 containing a held pair	853
`test_order`	test, order holdout	389

Held pair signatures (both orders each): (POL,PASS), (NUM,WH), (TENSE,CM), (POL,BIND); order holdout: (TENSE,NUM) with (NUM,TENSE) trainable; held triples: (NUM,TENSE,POL), (POL,NUM,PASS), (NUM,POL,WH), (POL,TENSE,BIND), none containing a held pair as an adjacent subword.

Table 2: Benchmark resources and model summary. Synthetic counts and model runs describe the OpClosure development instance (generator seed 17, validator passing). The operator lists record resource coverage, and Appendix D.1 specifies eligibility for interchange evaluation. The natural-language Family-A branch is final (round-3 inclusion mask applied): three annotation rounds completed, all five Family-A agreement dimensions confirmatory, and the round-3 inclusion mask applied to the model runs; Appendix I.6 explains the choice of the number–tense pair.

	Field	Value
synthetic	generator	typed operator grammar, deterministic realization, English-like
	operators	NUM, TENSE, POL, PASS, BIND, WH, CM (7 families)
	examples	3,701 (1,521 train / 493 val / 615 test / 1,072 lexical holdout); 2,880 are intervention bases
	intervention rows	108,362 across 17 named pools; training pool unary-only (6,802 rows)
	held-out policy	pairs held in both orders; triples free of held sub-pairs; order, depth-4, closure (3,084 rows), surface splits
	models	standard, shortcut, random, shuffled: seeds 0–4; positive control: seed 0 only (21 runs total)
	primary metric	held-out CIIA_3 (`test_triples_iid`)
natural lang.	language	English only
	operators	family A: cellwise-identifiable NUM, TENSE, POL (active-past NUM undefined; POL may span); family B: caused motion × verb semantics
	stimuli	minimal sets annotated in three rounds: family A has 128 targets plus 16 single-error controls; family B has 32 items. Round 3 passed both frozen family gates; its Family-A inclusion mask retains all 128 targets
	models	seven open-weight base LMs (360M–3B; revisions in Appendix D.2), weights frozen, float32
	primary metric	family A: TENSE×NUM, depth 2, both valid orders, equal-length candidate read-out; Family B documents the construction-transfer design

Split and evaluation checks. Class-B lexemes and held composition signatures are excluded from operator fitting and validation, and no held pair/triple surface transformation is reproduced by a train/validation intervention row. Operator fitting is unary-only; closure rows use test bases; and donors follow the same split discipline as their bases. The order-holdout is the intended exception to transformation exclusion because it isolates path dependence for a trainable high-level change. Evaluated interventions must also have an observable cue before the site and a distinct reflex after it; unobservable cases are excluded. All released instances pass these checks.

C.3.6 Effective-depth analysis of the development pools

This post-hoc descriptive analysis classifies every planned development-suite row key by effective depth (Definition 2, Table 3). The endpoint-genuine subset retains 37 of 103 base examples in the IID-triple pool and 5 of 65 in the depth-4 pool. 0.632 of closure rows are genuine.

The standard-model depth-3 minus unary contrast (D1) changes from -0.218 on all released rows to -0.427 on endpoint-genuine rows ([-0.538,-0.319]). The larger deficit on genuine rows shows that the easier reducible rows raised pooled accuracy. On endpoint-genuine IID triples, the positive-control minus standard contrast (D3) is +0.195 ([-0.016,+0.400]), with an interval that includes zero. The analysis retains each row's recorded score while changing its effective-depth grouping.

D Intervention and training details

This appendix connects the intervention procedure to the models, fitted maps, and controls used in evaluation. Appendix D.1 specifies sites, sources, and scoring, Appendix D.2 gives the training and fitting settings, and Appendix D.3 brings the validation gate together with the intervention and off-target controls.

D.1 Intervention protocol details

The synthetic protocol fixes the patching site, matches sources, and scores the resulting continuations. The development-suite selection and execution details below record the settings established before the independent OpClosure-v2 evaluation (Section 5.2).

Table 3: Effective-depth census of the released development-suite pools (post-hoc descriptive). A row is endpoint-genuine when no strictly shorter valid word reaches the same observable endpoint, so only a genuine depth-k row tests a depth-k composition. Right: the frozen contrast recomputed on the endpoint-genuine subset. The comparison groups the recorded row scores by effective depth.

Pool	bases	genuine bases	rows	genuine rows
`test_unary`	377	377	2500	2500
`test_pairs_iid`	259	259	2460	2460
`test_pair`	173	173	1325	1325
`test_order`	104	104	520	520
`test_triples_iid`	103	37	655	200
`test_triple`	101	101	530	530
`test_depth`	65	5	415	25
`test_closure`	269	241	2500	2035

	all released rows	endpoint-genuine rows	genuine bases
D1	-0.218 [-0.281, -0.157]	-0.427 [-0.538, -0.319]	37
D2	-0.082 [-0.104, -0.060]	-0.082 [-0.104, -0.060]	259
D4	+0.596 [+0.566, +0.626]	+0.596 [+0.565, +0.627]	259
D5	+0.641 [+0.613, +0.669]	+0.641 [+0.614, +0.669]	259
D6	+0.036 [-0.031, +0.100]	+0.036 [-0.032, +0.099]	108
D7	-0.201 [-0.230, -0.172]	-0.256 [-0.289, -0.225]	241

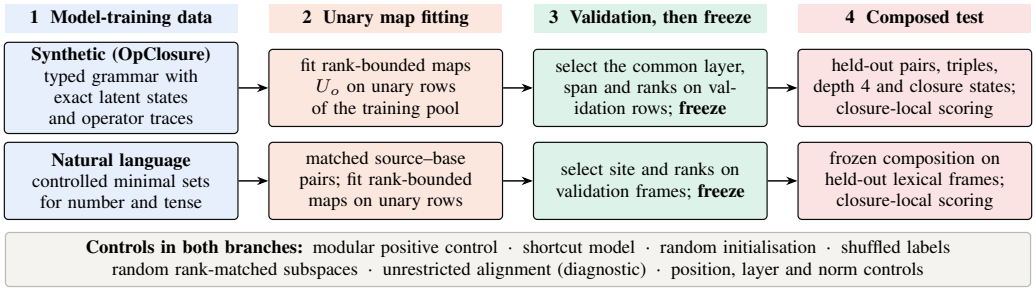


Figure 3: Benchmark and frozen evaluation pipeline. In both branches, rank-bounded maps are fitted on unary counterfactuals only (column 2), the site and ranks are selected on validation data and frozen (column 3), and the frozen maps are composed on held-out material, with each step scored closure-locally on the state it reaches (column 4). The synthetic branch uses the model conditions of Appendix D.2 and the controls of Table 6; the natural-language branch evaluates TENSE×NUM in both orders on held-out lexical frames (Appendix I).

D.1.1 SITES

The endpoint t^* is the echo-tail comma. At one Transformer block ℓ^*, $m \in \{1, 4, 8, 12\}$ is a nominal maximum look-back budget. The actual base block contains offsets $k = 0, \ldots, q_b - 1$, where $q_b = \min\{m, t_b^* + 1\}$; positions left of the first ordinary token are not intervened on. Base and source positions are right-aligned by offset from their respective endpoints. For a source with $q_s = \min\{m, t_s^* + 1\}$, offset $k < q_s$ uses its matching activation and $k \geq q_s$ reuses the leftmost available source activation at offset $q_s - 1$. One rank-r_o projector is tied across all actual base positions; $m = 1$ is retained as the single-position baseline. The block must end before the scored continuation so no counterfactual target token is placed inside the teacher-forced prefix.

The implementation follows this boundary rule, including clipping at the left boundary and source-edge replication when source and base widths differ. A separate check verifies the span semantics, high-level targets, and running-state source chains. The budget $m \sum_o r_o$ bounds the realized rank–site charge of each prefix.

The full validation grid is layer×span×rank. After excluding operators for which no cell passes the preregistered candidate-mass gate, the common (ℓ^*, m^*) maximizes the minimum validation unary CIIA over the remaining operators, with ties to smaller span and then earlier layer. At that common site, each operator takes the smallest rank within 0.01 CIIA of its own best. The resulting layer, span, ranks, exclusions, and all decision rules are frozen before the development-suite rerun.

Composition applies every operator sequentially at the selected common layer and span. The fresh development-suite validation-only sweep freezes layer 0/span 12, with ranks 1 for NUM and 8 for each of TENSE, POL, BIND, and PASS. CM fails the candidate-mass gate and is excluded before test-plan construction. The selected configuration passes the validation vacuity and candidate-mass checks.

D.1.2 SOURCE MATCHING

For operator o with target variable v_o, a source for base example b is any example whose latent state agrees with b's on every field except v_o and whose v_o equals the intervened value (for CM, the frame fields introduced by the operator are exempted from matching; for BIND, the source's theme is the rebind target). The generator precomputes up to three sources per (example, operator); evaluation uses the first. Sources always come from the same pool tier as training/evaluation rules allow (never from test into training).

D.1.3 OUTCOME SCORING

Given base prefix $x_{1:t^*}$ (through the site token), candidate set $\mathcal{C}$ (Appendix A.1), and composed operator $\Phi_w(\cdot; h_{s_{1:k}})$: run the model teacher-forced on $x_{1:t^*} \| c$ for each $c \in \mathcal{C}$ with the intervention applied across the span on every forward pass, sum continuation token log-probabilities including end-of-sequence, and renormalize over $\mathcal{C}$. CIIA compares the argmax to the generator's realization of $w(z_b)$; JSD compares the renormalized distribution to the one-hot target. The candidate set is a function of the base example alone (the complete finite fixed point of states reachable by valid words over the intervened operator family, deduplicated on realization), fixed before any intervention. Because renormalization can mask probability mass leaving the candidate set entirely, the *unnormalized* candidate mass $P(\mathcal{C}) = \sum_{c \in \mathcal{C}} p_\theta(c \mid \text{prefix}, \text{intervention})$ is reported alongside every CIIA cell, per condition and per depth; a CIIA gain accompanied by collapsing $P(\mathcal{C})$ is treated as a failure, not a success. JSD is evaluated only on each distribution's positive support, so exact softmax underflow cannot create a non-finite quantity; any remaining non-finite number aborts serialization.

D.1.4 FRAME-FIXING OPERATORS

Object WH questions lack the echo tail and therefore have no valid autoregressive intervention site; WH is exercised as a frame and behavioral stimulus only. PASS, by contrast, has a clause-structural cue before the comma and an auxiliary-family/subject reflex after it, so it is one of six autoregressive candidates alongside NUM, TENSE, POL, BIND, and CM. The development-suite validation gate excludes CM, leaving five active test operators. The counterfactual tables still cover all seven families.

D.1.5 CLOSURE EVALUATION

A closure row applies Φ_{o_2} directly to $h' = \Phi_{o_1}(h_b, h_{s_1})$, the activation produced by the first frozen intervention. Because o_2 acts on the state the first operator produced, its source must be matched to *that* state's frame rather than to the original base's; matching to the base is incorrect whenever o_1's variable belongs to o_2's frame, which is the usual case. Sources are therefore re-matched at analysis time against a train-split index, and closure rows for which no admissible observational source exists are counted and reported rather than scored—otherwise a closure failure could be an artefact of source mismatch rather than of composition. The development-suite evaluation recomputes availability from the frozen evaluation plan and reports it by pool and operator. The comparison target is $o_2(o_1(z_b))$ realized by the generator; scoring is as above with the composed function $h \mapsto \Phi_{o_2}(\Phi_{o_1}(h, h_{s_1}), h_{s_2})$. The generator supplies sampled closure rows, and the development protocol scores one frozen admissible source chain per eligible key: of $2{,}875$ rows involving the five active operators, 557 have a complete chain and the plan selects 500. Closure CIIA is therefore conditional on the declared eligible-row, two-stage distribution; unavailable rows and alternative admissible sources remain outside its estimand. The split targets the gap Proposition 1 predicts unary evaluation will miss.

D.1.6 FROZEN EVALUATION PLAN AND ESTIMAND

Before any development-suite organism is scored, a model-independent plan resolves every cheap eligibility condition: all word operators are active, a complete running-state-matched source chain and a nonempty boundary-clipped block exist for every base/source, and the target continuation exists and differs from the base. For each pool independently, SHA256-derived seeds order bases and order words within each base; round-robin traversal then applies the confirmatory cap of 500 scored rows. All organisms consume exactly the same frozen (base, word) keys, while traversal and exclusion counts retain their full uncapped denominators. Pool summaries average rows within base first and then average bases uniformly, matching $\mathcal{D}_k$ rather than weighting bases by how many valid words they admit.

D.2 MODELS AND TRAINING

Model training produces the synthetic model conditions. Operator fitting then learns the intervention maps with model weights fixed. The natural-language experiments start from pretrained weights (Section 5.3).

D.2.1 SYNTHETIC ORGANISMS

All synthetic organisms share one architecture: pre-LayerNorm decoder-only Transformer, a word-level vocabulary constructed from the released benchmark, learned positional embeddings, 4 blocks, $d_{\text{model}} = 128$, 4 heads, MLP ratio 4, and an untied output head; 830,464 parameters at the confirmatory configuration. The *standard* organism trains by next-token prediction (AdamW, learning rate $3 \cdot 10^{-4}$, batch 32, exactly 8,000 optimization steps; seedwise final cross-entropies 0.5281–0.5330) on the training partition of the observational corpus: all base surfaces plus train/validation-row counterfactual surfaces are partitioned by a fixed SHA256 hash of the full surface (19/20 train, 1/20 validation), so duplicate surfaces cannot cross the LM split. The language model sees the full lexicon (including class-B lexemes) so that lexical-holdout evaluation tests operator transfer rather than word knowledge, while operator fitting sees class-A unary rows only.

The *shortcut* organism trains for the same fixed budget on a semantic-state-filtered corpus omitting exactly the PRES+NEG quadrant. This creates the one-way support rule PRES⇒POS; it does not make polarity globally determined by tense.

The *random* organism is an untrained matched initialization. The *shuffled* organism uses standard weights and a fixed-point-free permutation π: fitting slot o consumes the complete unary rows, matched sources, and targets belonging to effective label $\pi(o)$. It does not relabel a row from o and synthesize a potentially domain-invalid $\pi(o)$ target.

The *positive control* is built by Interchange Intervention Training (Geiger et al., 2022): alongside the language-modeling loss, an auxiliary loss requires that copying designated disjoint coordinate blocks (one per operator) from a source activation into a base activation, with gradients flowing through both the donor encoder and the base path, at the oracle site layer 1 with nominal right-aligned maximum span 8 (using the same left-boundary clipping and source-edge extension), reproduces the unary counterfactual outcome. The auxiliary loss uses *unary training rows only*, so evaluating this organism on composed and closure pools remains a genuine test. Its role is to test whether the pipeline awards high closure-consistent CIIA to an organism trained to have a modular implementation. The positive control uses 6,000 fixed language-model steps and seed 0 only; the standard, shortcut, shuffled, and random organisms use seeds 0–4. Shortcut final cross-entropies are 0.5113–0.5182; shuffled runs reuse the corresponding standard weights and losses. The positive-control final language-model cross-entropy is 0.5667 (final IIT loss 0.00485), and the random organisms are untrained. Exact seedwise values and model hashes are retained in the primary JSON files.

D.2.2 OPERATOR FITTING

The implementation stores $W_o = U_o^\top$ (rank r, orthonormal rows via an orthogonal parametrization, so $P_o = W_o^\top W_o$) and trains it with Adam (learning rate 10^{-2}, exactly 300 optimization steps) to minimize cross-entropy of the candidate-set outcome under unary intervention against the generator's target, on at most 200 `train`-pool rows per operator. Model weights are frozen throughout fitting. Validation uses at most 80 rows per operator and sweeps all 4 layers, $m \in \{1, 4, 8, 12\}$, and

$r \in \{1, 2, 4, 8, 16, 32\}$. The frozen ordinary-model configuration has one common (ℓ^*, m^*) and a rank r_o^* per operator; the positive control is evaluated only at its oracle layer-1, span-8 coordinate interface.

D.2.3 NATURAL-LANGUAGE MODELS

The natural-language experiments use seven open-weight base models with frozen weights (Section 5.3). The protocol requires at least one family with public intermediate checkpoints and one independently developed family at two scales. The released configuration files record exact checkpoints, revisions, tokenizer versions, precision, hook implementations, sequence lengths, batch sizes, seeds, and hardware. The evaluated model revisions are Pythia-1.4B (EleutherAI/pythia-1.4b@fedc38a16eea); Pythia-2.8B (EleutherAI/pythia-2.8b@2a259cdd96a4); Pythia-410M (EleutherAI/pythia-410m@9879c9b5f8be); Qwen2.5-1.5B (Qwen/Qwen2.5-1.5B@8faed761d45a); SmolLM2-1.7B (HuggingFaceTB/SmolLM2-1.7B@effd688a1292); SmolLM2-360M (HuggingFaceTB/SmolLM2-360M@f8027fd0eaee); SmolLM3-3B (HuggingFaceTB/SmolLM3-3B@a07cc9a04f16) (final (round-3 inclusion mask applied)).

D.2.4 COMPUTE

Synthetic experiments run on CPU (the 21 confirmatory jobs completed in 166.5 minutes wall-clock with two concurrent two-thread workers on an Intel Core i7-13620H CPU). The reported natural-language runs used 14.3 GPU-hours on an NVIDIA GeForce RTX 4090 in float32, excluding snapshot downloads and gates.

D.3 VALIDATION AND INTERVENTION CONTROLS

Validation establishes whether a fitted map passes the declared gate before evaluation. The subsequent development-suite control analyses measure sensitivity to source choice, displacement magnitude, and patching location, as well as changes in non-target variables.

Before confirmation, every active operator must pass the frozen validation vacuity gate at its selected layer/span/rank. Its fitted unary CIIA minus the CIIA of a rank-matched random subspace must be strictly greater than 0.05; equality fails. Fitted, random-subspace, and observational candidate-mass quantities are computed on the same eligible validation rows and sites. The selected maps pass this gate: the fitted-minus-random-subspace CIIA gap is 0.5875 for TENSE and 1.0 for each other active operator, with candidate mass passing in every cell.

The auxiliary analyses probe source similarity, displacement magnitude, patching location, equal-value donors, and off-target effects (Tables 4–5). Surface alignment is summarized in Appendix J.1, and the map-family comparison appears in Table 30. Table 6 records the validation settings and analysis coverage.

These analyses use the development-suite standard models described in Appendix E.2.

Intervention and site controls. Table 4 reports the remaining controls of Appendix D.1 on the frozen unary rows. Replacing the fitted displacement by a norm-matched random direction leaves CIIA at most 0.01 and keeps the base continuation as argmax on at least 0.96 of rows, so the fitted effect is not a norm artefact. With a donor that equals the base on the operator's own variable but differs on another operator's variable, the frozen projector leaves the base continuation in place on 0.83–1.00 of rows and moves at most 0.19 renormalized mass onto the other operator's target, so the subspaces are not perfectly null on equal-value donors. Applying the layer-0 projector after block 1 or block 3 gives CIIA at most 0.57 and 0.00; the site offset alone gives at most 0.01 and the left-most actual offset alone at most 0.13, so the effect is carried by the distributed block rather than by any single position. Source-similarity correlations are listed in the same table.

Off-target budget. Table 5 maps the renormalized candidate distribution to the marginal of each high-level variable and compares intervened with observational marginals. Off-target movement is small for every operator: the largest off-target-to-on-target ratio is 0.02 for POL (maximum off-target JSD 0.018 bits against on-target 0.867), and non-target argmax flip rates range from 0.000 to

Table 4: Intervention and site controls on the frozen `test_unary` rows of the standard organism, seeds 0,1,2,3,4 (post-hoc descriptive). Models are reproduced as described in Appendix E.2. Ref: frozen intervention CIIA and target probability. Adjacent layers: the frozen layer-0 projector applied after block 1/2/3 with donor activations from that block (CIIA). Norm-matched random offset: the fitted displacement at every offset is replaced by a random direction of equal norm (CIIA, rate at which the base continuation stays argmax, candidate mass). No-op donor: the frozen projector applied with a donor that equals the base on the operator's own variable but differs on another active operator's variable, averaged over those donors (rate at which the base continuation stays argmax, i.e. the required null effect, and renormalized mass leaked onto the other operator's target). Site only / farthest only: CIIA when the frozen projector acts on a single offset, the site comma or the left-most actual offset of the block (first 30 rows per operator). Two-stage means.

	ref		adjacent layer			norm-matched random			no-op donor		single offset	
Operator	CIIA	P(tgt)	ℓ=1	ℓ=2	ℓ=3	CIIA	base	$P(\mathcal{C})$	base	leak	site	farthest
NUM	0.90	0.89	0.57	0.12	0.00	0.00	1.00	1.00	0.83	0.00	0.00	0.00
TENSE	0.79	0.78	0.01	0.00	0.00	0.01	0.99	1.00	0.89	0.16	0.01	0.00
POL	0.91	0.88	0.35	0.00	0.00	0.01	0.96	0.99	1.00	0.19	0.00	0.01
BIND	0.97	0.95	0.37	0.00	0.00	0.00	1.00	0.99	0.92	0.12	0.00	0.00
PASS	1.00	0.97	0.01	0.00	0.00	0.00	0.98	0.99	0.85	0.00	0.00	0.13

Source similarity (Spearman ρ across rows; $^*p < .05$): outcome vs. token Jaccard / activation distance / projected displacement

	vs. correct			vs. JSD to target		
Operator	Jaccard	act. dist.	proj. disp.	Jaccard	act. dist.	proj. disp.
NUM	$+0.19^*$	-0.17^*	-0.02	-0.66^*	$+0.63^*$	$+0.25^*$
TENSE	$+0.07$	-0.01	$+0.18^*$	-0.06	$+0.04$	-0.04
POL	$+0.13^*$	-0.08^*	-0.37^*	$+0.08^*$	-0.02	$+0.42^*$
BIND	$+0.04$	-0.10^*	-0.14^*	-0.25^*	$+0.34^*$	$+0.35^*$
PASS	–	–	–	-0.18^*	$+0.01$	-0.08

Table 5: Per-operator off-target (collateral) budget on the frozen `test_unary` rows of the standard organism, seeds 0,1,2,3,4 (post-hoc descriptive). Models are reproduced as described in Appendix E.2. The renormalized candidate distribution is mapped to the marginal of each high-level variable through the candidate→state map. On: JSD (bits) between the intervened and observational marginal of the operator's own variable and the intervened mass on its target value. Off: the same JSD for variables the operator does not write, reported as the per-row maximum and mean over observable non-target variables that admit more than one value on that base, the rate at which a non-target marginal's argmax flips away from the base value, and the off/on ratio (failure mode: off-target effects as large as target effects). Per-variable columns give the mean off-target JSD for that variable (– = that variable is the target or never observable). Two-stage means; n = bases.

		on-target		off-target				off-target JSD by variable				
Operator	n	JSD	mass	max JSD	mean JSD	flip	off/on	num	tense	pol	voice	theme
NUM	79	0.839	0.886	0.0006	0.0002	0.000	0.00	–	0.0006	0.0000	0.0001	0.0000
TENSE	146	0.737	0.779	0.0004	0.0001	0.000	0.00	0.0005	–	0.0000	0.0008	0.0000
POL	149	0.867	0.905	0.0183	0.0073	0.004	0.02	0.0293	0.0026	–	0.0014	0.0057
BIND	87	0.927	0.947	0.0000	0.0000	0.000	0.00	0.0000	0.0000	0.0000	0.0000	–
PASS	39	0.975	0.992	0.0091	0.0035	0.000	0.01	–	0.0085	0.0008	–	0.0011

0.004 across operators. Unary failures are therefore dominated by no-change outcomes (the base continuation keeps the argmax) rather than by changes of the wrong variable. Off-target effects as large as target effects are not observed at the marginal level on these rows.

E Statistical procedures and study versions

Paired comparisons average eligible rows within each base sentence and then compare conditions on shared bases. Appendix E.1 defines the comparison families and eligibility rules; Appendix E.2 summarizes the development and confirmatory datasets; and Appendix E.3 describes uncertainty across bases, training seeds, and benchmark instances.

Table 6: Validation selection and robustness coverage for the OpClosure development suite. Completed validation entries come from the sweep and selection artifacts. Post-hoc entries refer to descriptive auxiliary analyses performed after the confirmatory freeze, with intervention and off-target controls in Appendix D.3 and other analyses identified in the corresponding rows.

Axis	Levels committed	Status
layer ℓ	every block; one common layer selected	complete: $\ell^* = 0$
span m	$\{1, 4, 8, 12\}$; tied projector; one common span selected	complete: $m^* = 12$
rank r	$\{1, 2, 4, 8, 16, 32\}$	complete: NUM 1; TENSE/POL/BIND/PASS 8; CM excluded; exploratory test-set sweep: Table 29
source policy	running-state matched / no-op-equal-value donor	matched complete; no-op donor post-hoc: Table 4
organism controls	whole-label shuffled / random / shortcut / modular positive	complete
site controls	single-offset profile; adjacent layers; norm-matched offset	post-hoc: Table 4
order	both orders where valid; order-holdout paired analysis	raw pool complete; paired analysis post-hoc: Table 21
surface	operator-constant adverb perturbation (`test_surface`)	pool complete; alignment analysis: Appendix J.1
subsets	development-suite all-eligible rows	complete

E.1 Units, comparison families, and eligibility

E.1.1 Base-level means and paired comparisons

The unit of resampling is the *base example*: all intervention rows from one sentence enter or leave a bootstrap replicate together. Paired comparisons use the shared base set, averaging rows within base before averaging bases, matching the two-stage estimand D_k rather than a flat row mean.

E.1.2 OpClosure-v2 comparison family

OpClosure-v2 fixes six contrasts before the confirmatory instance is generated (Table 10). P1 compares closure-local with unary accuracy; P2 genuine depth-3 with unary; P3 and P4 standard with random and shuffled on IID pairs; P5 shortcut with standard; and P6 held with seen signatures. P1, P2, and P5 have negative target directions, P3 and P4 positive, while P6 is two-sided descriptive. The frozen procedure uses paired base-bootstrap intervals, sign-flip tests, and Benjamini–Hochberg correction at $q = 0.05$ over these six contrasts; confirmation also requires agreement with the hierarchical estimate, and P1 requires the declared sign on at least three of five seeds.

E.1.3 Development-suite comparison family

The development suite contains 3,701 examples and 108,362 intervention rows. Standard, shortcut, random, and shuffled conditions use seeds 0–4, while the modular reference uses seed 0. Its seven fixed contrasts are: D1, standard depth-3 versus unary; D2, shortcut versus standard; D3, modular reference versus standard at depth 3; D4/D5, standard versus random and shuffled; D6, held versus seen signatures; and D7, closure versus unary. Benjamini–Hochberg correction at $q = .05$ is applied over this seven-member family; other development-suite comparisons are descriptive.

Frozen contrast status. **D1**: $\Delta = -0.218$, 95% CI $[-0.280, -0.158]$, $p < .001$; confirmatory BH reject; **D2**: $\Delta = -0.082$, 95% CI $[-0.104, -0.060]$, $p < .001$; confirmatory BH reject; **D3**: $\Delta = +0.117$, 95% CI $[+0.020, +0.209]$, $p = 0.018$; confirmatory BH reject; **D4**: $\Delta = +0.596$, 95% CI $[+0.566, +0.627]$, $p < .001$; confirmatory BH reject; **D5**: $\Delta = +0.641$, 95% CI $[+0.613, +0.670]$, $p < .001$; confirmatory BH reject; **D6**: $\Delta = +0.036$, 95% CI $[-0.032, +0.099]$, $p = 0.293$; no confirmatory BH rejection; **D7**: $\Delta = -0.201$, 95% CI $[-0.230, -0.173]$, $p < .001$; confirmatory BH reject.

E.1.4 Eligibility and model runs

Eligibility is model-independent: each scored word must use active operators, have a complete running-state-matched source chain and valid span, and have a defined target distinct from the base. All organisms receive the same frozen keys within each pool. Runs failing the predeclared validation-loss criterion or a required candidate-mass gate cannot support a confirmatory result. In the completed development suite, none of the 16 trained runs is degenerate, all seven contrasts are eligible, and no included seed–pool cell fails the candidate-mass gate.

Table 7: Study versions. The development suite fixed the site, ranks, splits, and estimator; OpClosure-v2 is an independent confirmatory instance generated after the effective-depth correction. No development-suite row enters a v2 estimate, and the two suites are not pooled.

	OpClosure development suite	OpClosure-v2 (confirmatory)
generator	samples, then filters reducible words	filters by effective depth, then samples
generator seed	17	90901
training seeds	0–4 (0 for the positive control)	0–4 (0 for the positive control)
frozen family	seven contrasts (D1–D7)	six contrasts (P1–P6)
BH family size	$m = 7$	$m = 6$
depth pools	mixed genuine and endpoint-reducible	endpoint-genuine by construction
genuine depth-3 bases	37 of 103	105 rows, all genuine

Table 8: The frozen contrasts under a hierarchical bootstrap (post-hoc descriptive). The frozen procedure averages the five training seeds within a base example and then resamples bases, so its interval estimates base-sampling uncertainty only. The hierarchical procedure resamples training seeds and then bases, putting training-seed uncertainty inside the interval. Point estimates are identical by construction; only the intervals and the per-seed spread are new. $B = 10{,}000$.

	contrast	hierarchical 95% CI	per-seed range	covers 0
D1	standard IID triples vs unary	-0.218 [-0.281, -0.154]	-0.295 to -0.121	no
D2	shortcut vs standard, IID pairs	-0.082 [-0.168, +0.009]	-0.208 to +0.058	**yes**
D4	standard vs random, IID pairs	+0.596 [+0.480, +0.721]	+0.471 to +0.785	no
D5	standard vs shuffled, IID pairs	+0.641 [+0.523, +0.764]	+0.512 to +0.828	no
D6	held vs seen signature pairs	+0.036 [-0.137, +0.221]	-0.207 to +0.312	**yes**
D7	standard closure vs unary	-0.201 [-0.277, -0.125]	-0.292 to -0.084	no

E.2 Study versions

The development suite fixed the experimental settings before the independent OpClosure-v2 instance was generated. Table 7 summarizes the two versions. Auxiliary development analyses use the original models; Table 12 uses five independently retrained replication models, and Table 13 reports three further benchmark instances with their own confirmatory models.

E.3 Uncertainty across bases, training seeds, and instances

The frozen intervals average included seeds within each base and then resample bases (10^4 percentile-bootstrap replicates, with sign-flip tests), so they quantify base-sampling rather than training-seed uncertainty. The hierarchical analysis below resamples seeds and bases; crossed intervals in Table 12 use one shared base draw across resampled seeds, and the three-level summaries in Table 13 additionally resample benchmark instances.

Training-seed uncertainty. The frozen procedure averages the five training seeds within a base example and then resamples bases, so its intervals estimate base-sampling uncertainty and exclude the training-seed population. Resampling seeds first and then bases (Table 8) puts that uncertainty inside the interval. The point estimates are unchanged by construction. D1 and D7 survive ([-0.281,-0.154] and [-0.277,-0.125]), and their per-seed differences are one-signed throughout. D2, D6 cover zero under seed resampling. D2 is the substantive change: the shortcut organism sits -0.082 below standard on IID pairs with interval [-0.168,+0.009], and its per-seed differences run -0.208 to +0.058, positive on 1 of the five seeds. The shortcut model's lower IID-pair accuracy is supported by the base-level interval. The seed-resampling interval includes zero, reflecting variation in the ordering across training seeds. D6 covers zero under both estimators.

E.3.1 Shared-base correction

All five training seeds share the same base set. Drawing each resampled base cluster once and applying it to all selected seeds widens the recomputable hierarchical and three-level intervals by factors of 1.01–1.43 relative to independent within-seed base draws, without changing point estimates. For P1, only per-seed aggregates are available, so its reported hierarchical interval is unchanged.

Table 9: Graded realization profile. For each trained seed, α^* is the largest disagreement $1 - \mathrm{CIIA}$ over the pools a criterion includes: unary validation ($K{=}1$, `test_unary`) or the full criterion ($K{=}4$: unary, IID and held pairs, order holdout, IID and held triples, depth 4, closure). Entries are medians over seeds (the reference has one seed) and, after the slash, the number of seeds that pass at the development-fixed $\alpha_{\mathrm{dev}} = 0.40$, the smallest multiple of 0.05 at which the development reference passes $K{=}4$; passing also requires the candidate-mass gate in every included pool. Lower block: pretrained models on the number–tense pair (unary TENSE, unary NUM, both composition orders; seed-mean CIIA over three fitting seeds).

organism	criterion	dev.	conf.	repl.	90902	90903	90904
standard	$K{=}1$	0.11 / 5/5	0.22 / 5/5	0.18 / 4/5	0.16 / 5/5	0.15 / 5/5	0.10 / 5/5
	$K{=}4$	0.77 / 2/5	0.88 / 0/5	0.80 / 0/5	0.80 / 0/5	0.80 / 0/5	0.63 / 0/5
modular reference	$K{=}1$	0.02 / 1/1	0.02 / 1/1	0.01 / 1/1	0.06 / 1/1	0.05 / 1/1	0.12 / 1/1
	$K{=}4$	0.38 / 1/1	0.40 / 0/1	0.47 / 0/1	0.71 / 0/1	0.76 / 0/1	0.77 / 0/1
shortcut	$K{=}1$	0.21 / 5/5	0.27 / 3/5	0.34 / 3/5	0.29 / 5/5	0.26 / 5/5	0.20 / 4/5
	$K{=}4$	0.74 / 0/5	0.84 / 0/5	0.85 / 0/5	0.81 / 0/5	0.79 / 0/5	0.85 / 0/5
random	$K{=}1$	0.95 / 0/5	0.95 / 0/5	0.95 / 0/5	0.95 / 0/5	0.96 / 0/5	0.94 / 0/5
	$K{=}4$	0.97 / 0/5	0.97 / 0/5	0.97 / 0/5	0.96 / 0/5	0.97 / 0/5	0.98 / 0/5
shuffled	$K{=}1$	1.00 / 0/5	1.00 / 0/5	1.00 / 0/5	1.00 / 0/5	0.98 / 0/5	0.97 / 0/5
	$K{=}4$	1.00 / 0/5	1.00 / 0/5	1.00 / 0/5	1.00 / 0/5	1.00 / 0/5	1.00 / 0/5

pretrained models, number–tense pair: α^* over the four pools, pass at 0.20 / at α_{dev}
Pythia-410M 0.06 (yes/yes); Pythia-1.4B 0.08 (yes/yes); Pythia-2.8B 0.36 (no/yes); Pythia-6.9B 0.00 (yes/yes)
SmolLM2-360M 0.00 (yes/yes); SmolLM2-1.7B 0.12 (yes/yes); SmolLM3-3B 0.00 (yes/yes); Qwen2.5-1.5B 0.11 (yes/yes)

E.4 Graded realization profile

Definition 1 fixes a tolerance α before asking whether disagreement stays below it. Table 9 reports, for every trained seed, the smallest tolerance it meets, α^*, under two criteria: unary validation ($K{=}1$) and the full criterion over the eight declared pools up to depth four, closure states included ($K{=}4$). The operating point $\alpha_{\mathrm{dev}} = 0.40$ is fixed from the development suite alone: it is the smallest multiple of 0.05 at which the development modular reference meets $K{=}4$ (its α^* is 0.382, set by held pairs). Every other instance is classified at that tolerance.

Unary validation certifies the standard models on every instance (median α^* between 0.10 and 0.22), while the full criterion places them between 0.63 and 0.88: the tolerance a unary test licenses is not the tolerance composition reaches. Random and shuffled-label organisms fail both criteria everywhere. On the genuine-depth instances the modular reference also exceeds α_{dev} at $K{=}4$, bound by its depth-4 words. Seven of the eight pretrained models realize both operators and both composition orders of the number–tense pair at $\alpha^* \leq 0.125$, and so pass at $\alpha = 0.20$.

F Synthetic composition and local behavior

The synthetic results connect endpoint accuracy with the local behavior of reused operators. The original OpClosure-v2 comparisons are followed by matched analyses on replication models (Appendix F.2) and further benchmark instances (Appendix F.3). Appendices F.4 and F.5 examine the independent-error reference and source-policy comparison. The development-suite summaries appear in Appendix F.6. Interval procedures and the shared-base correction are given in Appendix E.3.

F.1 Confirmatory comparisons and local decomposition

Table 10 gives all six frozen comparisons on the original OpClosure-v2 instance, including the base-level and original hierarchical intervals and each member's status. The closure-local decomposition below resolves P1 by operator and reports its measurement coverage.

Closure-local decomposition. Table 11 evaluates the second operator against its high-level image on the state the model actually reached. This target follows the local comparison in assumption (A2) of Theorem 1. For the original confirmatory models, the pooled comparison pairs each base's closure and unary averages over the operators available in each pool. The paired closure-local minus

Table 10: OpClosure-v2 confirmatory family. The six contrasts, their required directions and both estimators were fixed before the confirmatory seed was generated. *frozen* resamples base examples with seeds averaged within base and decides, with sign-flip p and Benjamini–Hochberg at $q = 0.05$ over exactly these six members; *hierarchical* resamples seeds and then bases, so training-seed uncertainty sits inside its interval. A member whose two legs disagree is estimator-dependent, not confirmed; P1 also requires the sign on three of the five seeds. ***no (dir.)*** the interval does not meet the declared direction; *reported* no direction was declared.

Contrast	required	frozen (bases)	hierarchical (seeds, bases)	status
P1 closure-local	< 0	-0.242 [-0.280, -0.205] ($p < 0.001$)	-0.242 [-0.272, -0.212]	confirmed
P2 genuine depth 3 vs unary	< 0	-0.387 [-0.460, -0.313] ($p < 0.001$)	-0.387 [-0.480, -0.304]	confirmed
P3 standard vs random	> 0	+0.391 [+0.360, +0.422] ($p < 0.001$)	+0.391 [+0.296, +0.485]	confirmed
P4 standard vs shuffled	> 0	+0.418 [+0.388, +0.449] ($p < 0.001$)	+0.418 [+0.350, +0.481]	confirmed
P5 shortcut vs standard	< 0	+0.003 [-0.022, +0.027] ($p = 0.830$)	+0.003 [-0.072, +0.088]	**no (dir.)**
P6 held vs seen signature	reported, no sign	-0.130 [-0.181, -0.080] ($p < 0.001$)	-0.130 [-0.274, -0.019]	reported

Confirmed members: P1, P2, P3, P4.

Table 11: Closure-local decomposition (P1). The second operator is scored against its own image on the state the model actually reached, $o_2(\hat{\pi}(h_1))$, rather than against the ideal endpoint $o_2(o_1(z))$ — the quantity that assumption (A2) of Theorem 1 constrains. Δ is closure-local minus unary; unpaired compares pool means, paired compares the bases carrying both, frozen resamples those bases and hierarchical resamples training seeds and then bases. The operator rows pair the same operator on the same sentences; the pooled row pairs by sentence over whichever operators each pool applies to it, so it is a pool contrast and not a same-operator one. Rows whose reached state admits no decoded class, or at which o_2 is invalid, are unmeasurable and counted separately. † marks an operator whose frozen unary accuracy is below twice the random organism's; its row is kept in the pooled figure.

Operator	closure-local	unary	Δ unpaired	Δ frozen [95% CI]	Δ hier. [95% CI]
BIND	0.715 (93)	0.938 (80)	-0.222	-0.203 [-0.297, -0.114] (36)	-0.203 [-0.283, -0.133]
NUM	0.770 (79)	1.000 (64)	-0.230	-0.107 [-0.173, -0.053] (15)	-0.107 [-0.200, -0.027]
PASS	0.308 (24)	0.785 (39)	-0.476	—	—
POL	0.452 (130)	0.677 (155)	-0.225	-0.200 [-0.278, -0.122] (46)	-0.200 [-0.352, -0.070]
TENSE	0.383 (144)	0.663 (162)	-0.280	-0.281 [-0.357, -0.207] (54)	-0.281 [-0.394, -0.183]
pooled	0.515	0.759	—	-0.242 [-0.280, -0.205] (277)	-0.242 [-0.272, -0.212]

Coverage: 0 of 2500 scored closure rows unmeasurable (0.00%), of which 0 have no decoded class, 0 reach a state at which o_2 is invalid and 0 are quotient-ambiguous. Both steps and the endpoint correct: 1029. Error cancellation (first step wrong, second step wrong locally, endpoint nonetheless recovered): 57 rows, counted as endpoint recovery and never as compositional success.

unary gap is -0.242 (95% base-bootstrap CI [-0.280, -0.205] over 277 shared bases). The original hierarchical estimate is -0.242 ([-0.272, -0.212]), and the negative sign holds on 5 of the 5 training seeds.

The table also reports the same-operator comparisons and measurement coverage. The decomposition is unmeasurable on 0 of 2500 scored rows. Rows that recover the endpoint after two locally incorrect steps are recorded as endpoint recovery.

F.2 Matched comparisons on replication models

Table 12 reports descriptive row-level analyses of five cross-machine replication organisms trained under the frozen configuration (Appendix E.2). On these replication models, the P1 pool pairing gives -0.221 ([-0.254, -0.189]). This pooled value supplies the comparison on the same models used for the following views.

The strict comparison holds both the sentence and the second operator fixed. Its gap is -0.221 ([-0.262, -0.178] over 144 sentences and 156 sentence–operator cells), with crossed seed-and-base interval [-0.290, -0.155]. Restricting the closure rows to correctly decoded first steps gives -0.220 ([-0.268, -0.173]). On rows whose first step was incorrect, the gap is -0.305 ([-0.384, -0.225]). The deficit on the first subset identifies reduced local accuracy after a correct first step.

Table 12: Matched views of the closure-local contrast on replication models (descriptive). Rows come from five independently retrained replication models (Appendix E.2). *Pool* is the frozen P1 pairing (closure rows pooled by sentence against unary rows pooled by sentence); *strict* pairs the same second operator on the same sentence and pools over those cells, base-uniform; *first step correct/wrong* restricts the closure rows by whether the first operator decoded correctly; *state-matched* scores each closure row against the same operator's unary accuracy on corpus sentences whose latent state carries the reached state's profile (structural: voice, construction, gap; full: also number, tense, polarity), with the share of rows that have such a comparator; *decoded bases* keeps sentences the unintervened model decodes correctly on every seed. *frozen* resamples bases; *crossed* resamples seeds and then one base vector shared by all picked seeds.

View	Δ frozen [95% CI] (n)	Δ crossed [95% CI]	coverage
pool (P1 as frozen)	-0.221 [-0.254, -0.189] (277)	-0.220 [-0.273, -0.170] (277)	277 bases
strict, base-uniform	-0.221 [-0.262, -0.178] (144)	-0.221 [-0.290, -0.155] (144)	144 bases, 156 cells
strict, cell-level	-0.217 [-0.258, -0.178] (156)	—	same cells
first step correct, strict	-0.220 [-0.268, -0.173] (142)	—	142 bases, 154 cells
first step correct, pool	-0.225 [-0.259, -0.192] (276)	—	276 bases
first step wrong, strict	-0.305 [-0.384, -0.225] (103)	—	103 bases, 112 cells
first step wrong, pool	-0.239 [-0.295, -0.183] (213)	—	213 bases
state-matched, structural	-0.203 [-0.240, -0.167] (182)	—	63% of rows
state-matched, full	-0.217 [-0.271, -0.163] (139)	—	28% of rows; no reading
decoded bases, strict	-0.221 [-0.264, -0.178] (144)	-0.221 [-0.290, -0.156] (144)	367 bases
decoded bases, pool	-0.221 [-0.254, -0.189] (277)	-0.220 [-0.272, -0.170] (277)	base decoding 1.000 (closure), 1.000 (unary)
P2 genuine depth 3 vs unary, crossed	—	-0.387 [-0.494, -0.273] (85)	met, as the frozen leg
P3 standard vs random, crossed	—	+0.391 [+0.292, +0.488] (292)	met, as the frozen leg
P4 standard vs shuffled, crossed	—	+0.418 [+0.346, +0.487] (292)	met, as the frozen leg
P5 shortcut vs standard, crossed	—	+0.003 [-0.075, +0.094] (292)	not met, as the frozen leg
P6 held vs seen signature, crossed	—	-0.130 [-0.279, -0.003] (134)	no direction declared

Strict cells per operator: BIND 36, NUM 15, PASS 5, POL 46, TENSE 54. Crossed intervals are shown only where every seed scores the same base set.

State-matched comparisons ask how the same operator performs on observational sentences with the reached state's profile. Matching voice, construction, and gap gives -0.203 ([-0.240, -0.167]), with an available comparator for 63% of rows. Adding number, tense, and polarity to form the full six-field profile gives -0.217 ([-0.271, -0.163]), with coverage of 28%.

Conditioning on observational decoding addresses the model's accuracy before intervention. Before intervention, decoding accuracy is 1.000 on closure bases and 1.000 on unary bases. Among the 367 sentences decoded correctly on every seed, the strict comparison yields -0.221 ([-0.264, -0.178]). Crossed resampling changes none of the six-member family's direction verdicts.

F.3 FURTHER BENCHMARK INSTANCES

Table 13 reports three further benchmark instances, generated with the frozen generator under new seeds after every rule was frozen, each run end to end on one machine with its confirmatory checkpoints preserved.

The active operator set is whatever the frozen candidate-mass gate left in each instance: 90902 five active operators (CM excluded by the candidate-mass gate); 90903 five active operators (CM excluded by the candidate-mass gate); 90904 six active operators (none excluded). Its auxiliary analysis uses the six-operator scripts. Each instance resolves the six-member confirmatory family by the v2 rules on its own; under the frozen base-level reading with BH correction the members are confirmed by P1 3 of 3, P2 3 of 3, P3 3 of 3, P4 3 of 3, P5 3 of 3, P6 0 of 3 instances, with P5 (shortcut against standard) resolved as 90902 **estimator-dependent**; 90903 **estimator-dependent**; 90904 confirmed and P6 as 90902 reported; 90903 reported; 90904 reported.

On the confirmatory organisms themselves, the same-sentence same-operator closure-local deficit lies between -0.170 and -0.258 across instances (three-level interval [-0.264, -0.157], resampling instances, seeds and bases), the pool contrast's three-level interval is [-0.235, -0.165], and the genuine depth-3 degradation lies between -0.351 and -0.418 (three-level [-0.467, -0.311]).

The IIT-trained modular positive control, scored at its oracle site under the same metric, has unary accuracy between 0.877 and 0.950, closure-local accuracy between 0.747 and 0.806, and a same-

Table 13: Three further OpClosure-v2 benchmark instances generated with the frozen generator under new seeds, each run end to end on one machine with its confirmatory checkpoints preserved, so that the same-sentence same-operator contrast (strict) is computed on the confirmatory organisms themselves. Each instance resolves the six-member confirmatory family by the v2 rules; the seed-90901 instance of Table 10 is shown for reference, its strict value coming from the cross-machine replication organisms. *PC gap* is the same-sentence same-operator closure-local gap of the IIT-trained modular positive control at its oracle site; *std gap* the standard organisms' five-seed mean of the same quantity. Intervals are base-level paired bootstraps; the three-level interval resamples instances, then seeds, then one base vector per instance.

instance	P1 pool	P1 strict	P2 genuine depth 3	confirmed	PC gap (unary, closure-local)	std gap
90901 (v2)	-0.242 [-0.280, -0.205]	-0.221 [-0.262, -0.178]†	-0.387 [-0.460, -0.313]	P1, P2, P3, P4	—	—
90902	-0.203 [-0.236, -0.170]	-0.204 [-0.245, -0.164] (136)	-0.418 [-0.474, -0.362]	P1, P2, P3, P4	-0.191 [-0.257, -0.125] (0.940, 0.780)	-0.204
90903	-0.178 [-0.211, -0.146]	-0.170 [-0.210, -0.132] (137)	-0.351 [-0.399, -0.304]	P1, P2, P3, P4	-0.111 [-0.175, -0.047] (0.950, 0.806)	-0.169
90904	-0.220 [-0.252, -0.187]	-0.258 [-0.306, -0.212] (135)	-0.402 [-0.471, -0.333]	P1, P2, P3, P4, P5	-0.150 [-0.213, -0.091] (0.877, 0.747)	-0.258
three-level, P1 pool	-0.199 [-0.235, -0.165]	instances -0.201, -0.175, -0.221				
three-level, P1 strict	-0.209 [-0.264, -0.157]	instances -0.204, -0.164, -0.258				
three-level, P2	-0.390 [-0.467, -0.311]	instances -0.418, -0.351, -0.402				

Instances confirming each member under the frozen base-level reading with BH correction: P1 3/3, P2 3/3, P3 3/3, P4 3/3, P5 3/3, P6 0/3. Status by instance for the members that are not confirmed outright: P5 90902 **estimator-dependent**; 90903 **estimator-dependent**; 90904 confirmed; P6 90902 reported; 90903 reported; 90904 reported. † replication organisms retrained under the frozen configuration (Appendix E.2).
Active operators by instance, as the frozen candidate-mass gate resolved them: 90902 five active operators (CM excluded by the candidate-mass gate); 90903 five active operators (CM excluded by the candidate-mass gate); 90904 six active operators (none excluded). Its auxiliary analysis uses the six-operator scripts.

sentence same-operator gap between -0.111 and -0.191, against -0.169 to -0.258 for the standard organisms of the same instances.

F.4 An independent-error reference for depth accuracy

This descriptive analysis compares the observed depth-3 minus unary contrast with an independent-per-step reference. If each of three steps succeeds independently with unary accuracy u, endpoint accuracy is u^3. On the same 85 bases in the confirmatory instance, $u = 0.775$ and the observed contrast is -0.387.

The reference predicts a contrast of -0.309. Under this reference, ordinary per-step accumulation accounts for about 80% of the depth-3 drop, leaving a composition-specific excess of -0.078. Closure-local scoring measures the next operator's action on the state it receives (Section 4.3).

F.5 Source matching along a composition

This descriptive comparison tests whether the closure-local deficit depends on the source policy. The running policy matches donors to the running high-level state obtained by the preceding high-level operations. The alternative matches each step to the original base. Both policies give a negative closure-local gap in the replication set and instances 90902 and 90903, so the deficit is not an artifact of the source policy.

set	running-policy gap	base-policy gap
replication set	-0.221	-0.207
instance 90902	-0.203	-0.158
instance 90903	-0.178	-0.176

F.6 Development-suite summaries

Table 14 and Figure 4 summarize the validated OpClosure development suite. Its frozen contrasts and statistical status are recorded in Appendix E.1, and Table 7 identifies the artifacts for the development suite and the independent OpClosure-v2 suite. The validated development-suite aggregate retains the per-seed summaries behind the organism-level means. Off-target diagnostics and the probing, direct, and unrestricted baseline families of the map-family comparison appear in Tables 5 and 30.

Table 14: OpClosure development results from the strictly validated suite. Entries are organism-level means over n_s included seeds for tail-token behavior, CIIA by word length, lexical transfer, unary candidate mass, and the selected site with nominal rank–site budget $m \sum_o r_o$. Appendix F.6 describes the per-seed summaries and sources. Statistical intervals, off-target diagnostics, and baseline families of the map-family comparison appear in Appendix E.3 and Tables 5 and 30.

Organism	Behavior	CIIA_1	CIIA_2	CIIA_3	CIIA_4	Lex	$P(\mathcal{C})$@1	site	budget
positive control ($n_s = 1$)	0.996	0.980	0.842	0.799	0.831	0.870	0.771	ℓ=1, m=8	64
standard Transformer ($n_s = 5$)	0.996	0.888	0.644	0.681	0.636	0.607	0.949	ℓ=0, m=12	396
shortcut model ($n_s = 5$)	0.984	0.776	0.562	0.578	0.547	0.512	0.921	ℓ=0, m=12	396
random model ($n_s = 5$)	0.008	0.053	0.048	0.048	0.078	0.073	0.000	ℓ=0, m=12	396
shuffled-label align. ($n_s = 5$)	0.996	0.007	0.003	0.132	0.026	0.022	0.984	ℓ=0, m=12	396

Table 15: Closure-local decomposition on the development suite (post-hoc descriptive). The second operator is scored against its image on the state the model actually reached rather than against the ideal endpoint, which is the quantity the composition theorem's uniformity assumption constrains. Δ is closure-local minus the same operator's accuracy on matched observational rows.

Operator	closure-local	unary	Δ	Δ paired [95% CI]
BIND	0.734	0.966	-0.231	-0.300 [-0.482, -0.136] (22)
NUM	0.854	0.901	-0.047	-0.125 [-0.208, -0.058] (24)
PASS	0.748	1.000	-0.252	—
POL	0.786	0.907	-0.121	-0.103 [-0.152, -0.055] (33)
TENSE	0.588	0.788	-0.200	-0.200 [-0.268, -0.137] (41)
pooled	0.739	0.888	—	-0.156 [-0.186, -0.125] (265)

1 of 2500 scored closure rows is unmeasurable. Error cancellation (first step wrong, second step wrong locally, endpoint recovered) occurs on 49 rows and is counted as endpoint recovery, not as compositional success.

All trained development-suite seeds passed the degeneration gate, and all included seed–pool cells passed the candidate-mass gate. Appendix D.1 reports the selected configuration from the full layer–span–rank validation grid. The local decomposition below reports the development suite's post-hoc measurements, and Figure 4 shows its descriptive depth, ordered-pair, and synthetic-transfer views.

Closure-local accuracy on the development suite. The development-suite post-hoc descriptive decomposition scores the second operator against its high-level image on the reached state (Table 15). The pooled paired difference from accuracy on matched observational rows is -0.156 ([-0.186,-0.125] over 265 shared bases). Among the unpaired operator-level gaps, PASS has the largest deficit at -0.252. This local loss shows that composed decline includes errors in the reused operator's own action on reached states, alongside propagation of earlier errors.

The decomposition is unmeasurable on 1 of 2500 scored rows. 49 rows recover the endpoint after both steps were locally incorrect. These rows are recorded as endpoint recovery, with their local failures retained. The corresponding OpClosure-v2 analysis is reported in Table 11.

G Controlled references and training

Controlled references calibrate the measured closure-local deficit and identify the model-attributable excess. Appendix G.1 reports the unary-trained modular reference on the replication set. Appendix G.2 compares fitted and designated-block interchanges and introduces the constructed stale-copy readout, with the paired excess and its pooled summary reported together. Appendix G.4 gives the six-member family trained with correct or stale composed targets.

G.1 The unary-trained modular reference

Unary-trained modular reference on the replication set. On the replication set, the unary-trained modular reference is evaluated with the protocol's fitted maps at its designated coordinate interface, layer 1 with nominal span 8. Its auxiliary training loss targets unary counterfactuals under designated-block interchanges (Appendix D.2). This evaluation measures how the fitted operators act after a first intervention in a model trained with that interface.

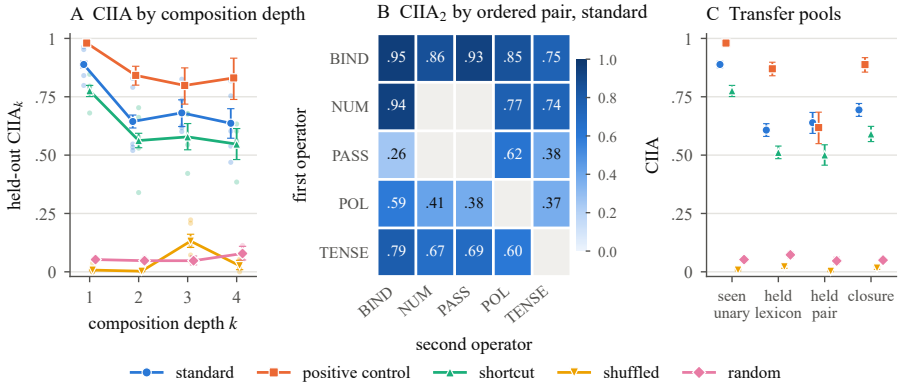


Figure 4: OpClosure development summaries. Panel A shows CIIA by word length in the original development pools, whose effective-depth composition is reported in Appendix C.3.6. Panel B shows ordered-pair accuracy for the standard organism, and panel C compares seen unary, lexical-holdout, held-pair, and closure pools. These views are descriptive. Points in A show individual seeds. Error bars are 95% base-bootstrap intervals after averaging seeds and quantify base-sampling uncertainty.

The IIT-trained modular positive control, reproduced on the same machine and scored under the same decomposition at its oracle site (layer 1, span 8), has unary accuracy 0.995, closure-local accuracy 0.830, closure endpoint accuracy 0.831, genuine depth-3 accuracy 0.833, and a same-sentence same-operator gap of -0.177 [-0.240, -0.118] over 144 sentences, with 1 error-cancellation rows; the standard organisms of the same set average 0.753 unary, 0.534 closure-local and a gap of -0.221.

The modular reference's designated-block mechanism composes exactly under its own interventions. Under fitted maps, both the reference and the standard models are less accurate for the same operator on reached states than on matched base sentences, with a smaller gap for the reference. Part of the measured deficit thus travels with the intervention protocol itself. Table 16 extends this calibration to oracle designated-block interchanges and a constructed stale-copy readout across four organism sets.

G.2 Readout construction and paired excess

Reference conditions for the closure-local diagnostic. Table 16 compares the standard models with a unary-trained modular reference on four organism sets, comprising the replication set and three further benchmark instances. The reference's designated-block mechanism composes exactly under its own interventions. At its designated site, layer 1 with nominal span 8, we evaluate both oracle designated-block interchanges and the protocol's fitted rank-width maps. Mean closure-local accuracies are 0.784 and 0.791, respectively. The standard-model rows report five-seed means under fitted maps.

The stale-copy condition changes the readout of the trained reference model. After each operator, it restores every other designated block to its pre-intervention value. This construction adapts the stale-copy idea of Proposition 1 to the reference model's designated blocks. It passes the declared unary check on every set, with unary accuracy within 0.02 of the reference under the same maps. Under fitted maps, its mean closure-local accuracy is 0.060 and its mean gap is -0.861. Under oracle maps, its mean gap is -0.850. It supplies the stale-copy end of the calibrated scale while preserving similar unary accuracy.

The accuracy columns average the frozen plan rows with equal weight per base. The strict gap pairs closure-local and unary outcomes for the same operator on the same base sentence. The model-attributable excess subtracts the reference's fitted-map gap from the standard models' gap, pairing the two by base sentence. Its mean is -0.056, and it ranges from -0.108 to -0.013 across sets. The

Table 16: Reference conditions for closure-local behavior. The positive control is the unary-trained modular reference. At its designated site, layer 1 with nominal span 8, *oracle* maps exchange designated blocks and *fitted* maps use the protocol's rank-width fits. The stale-copy control changes this model's readout to restore every other designated block to its pre-intervention value after each operator, adapting the construction of Proposition 1. *Unary*, *closure-local*, and *endpoint* are base-uniform accuracies on the frozen plan rows. The *gap* pairs closure-local minus unary outcomes for the same sentence and operator. Intervals use a paired base bootstrap. Standard rows report five-seed means under fitted maps. *Excess* is the standard models' gap minus the reference's fitted gap, paired by sentence. *Share* is the reference's fitted gap minus its oracle gap. *Position* divides the excess by the difference between the fitted stale-copy gap and the fitted reference gap. These comparisons are descriptive.

set	organism, maps	unary	closure-local	endpoint	gap
replication set	standard, fitted (seed mean)	0.753	0.534	0.447	-0.221
	positive control, fitted	0.995	0.830	0.831	-0.177 [-0.240, -0.122]
	positive control, oracle	0.986	0.847	0.853	-0.156 [-0.219, -0.097]
	stale-copy control, fitted	0.992	0.010	0.010	-0.972 [-0.993, -0.944]
	stale-copy control, oracle	0.986	0.010	0.010	-0.965 [-0.990, -0.934]
instance 90902	standard, fitted (seed mean)	0.837	0.640	0.583	-0.204
	positive control, fitted	0.940	0.780	0.775	-0.191 [-0.257, -0.125]
	positive control, oracle	0.886	0.743	0.705	-0.169 [-0.232, -0.110]
	stale-copy control, fitted	0.945	0.081	0.054	-0.827 [-0.890, -0.757]
	stale-copy control, oracle	0.886	0.057	0.005	-0.803 [-0.869, -0.733]
instance 90903	standard, fitted (seed mean)	0.872	0.697	0.634	-0.169
	positive control, fitted	0.950	0.806	0.770	-0.111 [-0.175, -0.049]
	positive control, oracle	0.941	0.809	0.775	-0.095 [-0.157, -0.035]
	stale-copy control, fitted	0.944	0.052	0.014	-0.856 [-0.911, -0.794]
	stale-copy control, oracle	0.941	0.059	0.021	-0.839 [-0.898, -0.776]
instance 90904	standard, fitted (seed mean)	0.911	0.696	0.672	-0.258
	positive control, fitted	0.877	0.747	0.675	-0.150 [-0.215, -0.091]
	positive control, oracle	0.860	0.739	0.666	-0.202 [-0.272, -0.131]
	stale-copy control, fitted	0.873	0.096	0.017	-0.789 [-0.859, -0.715]
	stale-copy control, oracle	0.860	0.083	0.011	-0.793 [-0.859, -0.719]
replication set	excess -0.044 [-0.101, +0.015]; share -0.021 [-0.076, +0.031]; position 0.06				
instance 90902	excess -0.013 [-0.082, +0.055]; share -0.022 [-0.081, +0.037]; position 0.02				
instance 90903	excess -0.059 [-0.119, +0.000]; share -0.016 [-0.057, +0.022]; position 0.08				
instance 90904	excess -0.108 [-0.169, -0.046]; share +0.052 [+0.000, +0.107]; position 0.17				

Across the four sets the excess averages -0.056 (from -0.108 to -0.013; its interval excludes zero on 1 of 4 sets) and the share averages -0.002. The stale-copy control is a valid witness on every set (unary within 0.02 of the positive control's under the same maps).

per-set paired base-bootstrap intervals exclude zero in 1 of 4 sets. This descriptive comparison identifies the part attributable to ordinary training beyond the modular reference.

For the same modular reference, the fitted-minus-oracle gap, labeled *share*, averages -0.002. This isolates the fitting protocol's contribution on the same designated-block mechanism. The normalized *position* divides the paired model-attributable excess by the fitted stale-copy gap minus the fitted reference gap. Its reported mean is 0.08, locating the standard models between the modular and stale-copy ends.

G.2.1 Pooled model-attributable excess

The paired model-attributable excess subtracts the unary-trained modular reference's fitted-map gap from the standard models' strict closure-local minus unary gap. The comparison pairs the two gaps by base sentence. Table 16 gives this comparison for each of the four organism sets. The per-set intervals are the primary summaries, with one of four excluding zero. The excess is negative in all four sets, giving a one-sided sign-test $p = 0.0625$.

A descriptive pooled analysis combines the 552 bases with equal weight for each set. The mean model-attributable excess is -0.056 with interval $[-0.087, -0.025]$ and sign-flip $p = 0.0004$. Treating the four sets as bootstrap clusters gives an interval of $[-0.102, -0.010]$.

Table 17: Closure-local minus unary gap after donor realignment, on the replication models and two further instances (depth-2 closure rows; cell = base sentence × second operator, $o_1 \neq o_2$). *Frozen*: second donor right-aligned as in the executed protocol. *Realigned*: second donor aligned by grammatical role, everything else frozen. Excess: standard minus modular reference on the same cells. Intervals: 10,000 base resamples per set; pooled rows weight the three sets equally. The last column is a t interval over the three set estimates. Estimators fixed before computation (Appendix G.3).

	frozen, pooled	realigned, pooled	realigned, replication	realigned, 90902	realigned, 90903	instance-level
standard organisms (5 seeds)	−0.189 [−0.214, −0.166]	−0.124 [−0.145, −0.105]	−0.156 [−0.193, −0.120]	−0.108 [−0.145, −0.073]	−0.109 [−0.143, −0.076]	[−0.192, −0.057]
modular reference, fitted maps	−0.140 [−0.177, −0.104]	−0.048 [−0.075, −0.024]	−0.050 [−0.093, −0.014]	−0.069 [−0.119, −0.023]	−0.026 [−0.075, +0.022]	[−0.102, +0.005]
modular reference, exact interchanges	−0.125 [−0.160, −0.091]	−0.035 [−0.061, −0.010]	−0.021 [−0.057, +0.014]	−0.069 [−0.115, −0.027]	−0.015 [−0.063, +0.034]	[−0.109, +0.038]
excess over fitted-map reference	−0.050 [−0.086, −0.014]	−0.076 [−0.105, −0.047]	−0.106 [−0.151, −0.058]	−0.038 [−0.092, +0.015]	−0.083 [−0.133, −0.034]	[−0.161, +0.009]
excess over exact-interchange reference	−0.064 [−0.098, −0.030]	−0.089 [−0.116, −0.062]	−0.134 [−0.178, −0.091]	−0.038 [−0.085, +0.011]	−0.094 [−0.145, −0.046]	[−0.209, +0.031]

G.3 Donor realignment and the paired excess

The frozen protocol right-aligns each donor block from the site comma. When the first operator changes the surface length at the site (POL or PASS), the second donor's cues therefore land at offsets that unary rows never use. Table 17 rescores the depth-2 closure rows of the replication models and of instances 90902 and 90903 with the second donor aligned by grammatical role and everything else frozen: models, maps, first steps, candidates and targets. A cell is a base sentence and a second operator with $o_1 \neq o_2$, a measurable closure row and a frozen unary row of the same operator on the same base. The standard organisms average their five seeds within the cell, and the excess subtracts the modular reference's gap on the same cells.

Realignment removes most of the reference's gap and less of the standard models'. The reference moves from -0.140 to -0.048 [-0.075, -0.024], and the standard models from -0.189 to -0.124 [-0.145, -0.105]. The paired excess therefore grows from -0.050 to -0.076 [-0.105, -0.047] against fitted maps, and to -0.089 [-0.116, -0.062] against exact designated-block interchanges; it is negative on every set. The analysis uses the same cell-level pairing and resampling procedure across all three datasets. We restrict it to the instances whose refits reproduce on the evaluation machine.

G.4 Supervision with composed interchanges

Closure training removes the deficit and stale training reproduces it. The six trained models in Table 18 extend the modular reference's recipe with composed interchange rows. For member k, the stale set S_k contains the first k variables in the order NUM, TENSE, POL, BIND, PASS. When two designated blocks have changed, the composed target restores each variable in S_k to its base value. Member 0 has an empty stale set and uses the correct high-level composition target. All members use the replication dataset and model-training seed 0. Evaluation uses either the protocol's fitted rank-width maps or oracle designated-block interchanges.

Under fitted maps on the same replication dataset, adding correct composed interchange targets raises closure-local accuracy from 0.830 for the unary-trained modular reference to 0.942 for member 0. It is the strongest organism under the closure-local metric among this trained family, the unary-trained reference, and the standard models. Its unary and closure endpoint accuracies are 0.999 and 0.936, respectively. Training on reached states closes most of the gap, with the strict closure-local minus unary contrast changing from -0.177 to -0.087 (95% paired base-bootstrap CI [-0.132, -0.045]). The standard-model gap is -0.221. The accuracy columns summarize their respective pools, while the strict gap pairs the same base sentence and operator across pools.

For each member, an affected row has its first or second operator in S_k. The expected stale share is the fraction of measurable closure rows meeting this condition. Member 1 designates NUM as stale. It retains unary accuracy 0.959 and loses closure-local accuracy exactly where it was taught to, scoring 0.769 on affected rows and 0.952 on other rows. For member 5, the reported closure-local accuracy is 0.315 and the strict gap is -0.205 ([-0.274, -0.142]). Table 18 retains the full row groups and results under both fitted and oracle maps.

Members 0 and 1 reach fitted-map unary accuracy of at least 0.8; members 2 to 5 reach 0.797, 0.554, 0.640, 0.510, respectively. Across all six members, the Spearman correlations between expected

Table 18: Model training with correct and stale composed targets. Member k adds composed interchange rows to the modular reference's training recipe. Its targets restore the variables in S_k to their base values, with S_k containing the first k of NUM, TENSE, POL, BIND, PASS. Member 0 uses the correct high-level composition target. A row is affected when its first or second operator lies in S_k. *Expected* is the fraction of measurable closure rows that are affected. *Closure-local* reports accuracy over all rows, followed by accuracies on affected and other rows in parentheses. The strict *gap* pairs closure-local minus unary accuracy for the same base sentence and operator. Displayed gap intervals are 95% paired base-bootstrap intervals. Fitted maps are the protocol's rank-width refits, and oracle maps are designated-block interchanges. The trained members use the replication dataset and model-training seed 0. The positive control is the unary-trained modular reference, and the standard row reports a five-seed mean.

member k	stale set	expected	unary	closure-local (stale / fresh rows)	gap, fitted	closure-local, oracle	gap, oracle
0	none	0.00	0.999	0.942 (— / 0.940)	-0.087 [-0.132, -0.045]	0.937	-0.087 [-0.132, -0.045]
1	NUM	0.34	0.959	0.886 (0.769 / 0.952)	-0.066 [-0.108, -0.031]	0.841	-0.062 [-0.104, -0.028]
$2^{\ddagger}$	NUM, TENSE	0.79	0.797	0.597 (0.549 / 0.864)	-0.181 [-0.243, -0.122]	0.515	-0.149 [-0.208, -0.097]
$3^{\ddagger}$	NUM, POL, TENSE	0.96	0.554	0.357 (0.367 / 0.444)	-0.185 [-0.262, -0.112]	0.279	-0.070 [-0.129, -0.014]
$4^{\ddagger}$	BIND, NUM, POL, TENSE	1.00	0.640	0.392 (0.398 / —)	-0.247 [-0.316, -0.180]	0.235	-0.090 [-0.146, -0.038]
$5^{\ddagger}$	BIND, NUM, PASS, POL, TENSE	1.00	0.510	0.315 (0.322 / —)	-0.205 [-0.274, -0.142]	0.168	-0.167 [-0.233, -0.104]
positive control	—	0.00	0.995	0.830	-0.177 [-0.240, -0.122]	0.847	-0.156 [-0.219, -0.097]
standard (seed mean)	—	—	0.753	0.534	-0.221	—	—

Spearman correlation between the expected stale share and the measured closure-local loss: over all 6 members, descriptively, +0.90 under the fitted maps and +0.99 under the oracle maps. ‡ marks a member with unary accuracy below 0.8.

stale share and closure-local loss are +0.90 under fitted maps and +0.99 under oracle maps. These all-member correlations are descriptive.

The composed training targets supervise sampled interchanges, and closure-local evaluation measures the resulting local behavior on the declared evaluation rows. Theorem 1 bounds composition error using uniform one-step control throughout the intervention closure.

H Generalization and explanation classes

The analyses below examine variation across evaluation pools and explanation classes. Appendix H.1 reports sequence, lexical, and order comparisons in the OpClosure development suite; Appendix H.2 compares explanation classes on the replication models.

H.1 Sequence, lexical, and order comparisons

These descriptive development-suite views resolve the pooled summaries in Appendix F.6. The full pool table reports sequence and lexical holdouts, the ordered-pair matrices separate operator signatures, and the order comparison pairs two paths to the same high-level target.

All evaluated pools. Table 19 lists every one of the ten frozen pools for every organism, because the fixed layout of Table 14 and Figure 4 shows only a subset. Two pools that no frozen contrast covers deserve explicit mention. On held-signature triples (`test_triple`; 101 bases, all words with three distinct operators) the standard organism reaches $\mathrm{CIIA}_3 = 0.370$ (95% CI [0.329,0.412]) against 0.681 ([0.623,0.740]) on the IID-triple pool. That pool gap is a word-structure effect, not a held-signature penalty: the IID pool mixes 29 bases whose word has three distinct operators, which score 0.366—the same as the held triples—with 86 bases whose word repeats an operator (two applications of a toggle cancel in the high-level target), which score 0.769. The shortcut organism shows the same pattern (0.277 held versus 0.372 seen-distinct), whereas the modular positive control stays at 0.792 against 0.828. Three distinct toggles are therefore hard for the standard organism whether or not their signature was seen, and the pooled CIIA_3 entering D1 is buoyed by repeated-operator words. These triple-pool comparisons are descriptive. On the surface-perturbed pool the standard organism reaches 0.433 against 0.888 on unperturbed unary rows. Appendix J.1 shows that role-based donor realignment recovers most of this surface perturbation loss.

Ordered-pair matrices. Table 20 gives the full ordered-pair composition matrix for every organism, with signatures that have no eligible frozen row masked rather than zeroed. Within the standard

Table 19: All ten frozen pools for every organism (descriptive display of the validated development-suite artifacts; the fixed Table 14 layout shows only a subset). Each entry is the two-stage CIIA over the frozen evaluation-plan rows with a percentile bootstrap 95% interval over bases after seed averaging (10,000 replicates; training-seed uncertainty excluded); n = bases, shared by all organisms. Values are the organism means in `aggregate.json`. Appendix E.1 specifies the seven confirmatory contrasts. The held-signature triple and surface-perturbed pools are descriptive comparisons.

Pool	n	positive control	standard	shortcut	shuffled	random
unary	377	0.980 [0.97,0.99]	0.888 [0.87,0.90]	0.776 [0.75,0.80]	0.007 [0.00,0.01]	0.053 [0.04,0.06]
IID pairs	259	0.842 [0.80,0.88]	0.644 [0.62,0.67]	0.562 [0.53,0.59]	0.003 [0.00,0.01]	0.048 [0.04,0.06]
held pairs	173	0.618 [0.55,0.68]	0.639 [0.59,0.68]	0.501 [0.45,0.54]	0.002 [0.00,0.00]	0.047 [0.03,0.06]
held order	104	1.000 [1.00,1.00]	0.669 [0.64,0.70]	0.475 [0.42,0.53]	0.000 [0.00,0.00]	0.063 [0.04,0.09]
IID triples	103	0.799 [0.72,0.87]	0.681 [0.62,0.74]	0.578 [0.52,0.63]	0.132 [0.10,0.16]	0.048 [0.03,0.07]
held triples	101	0.792 [0.71,0.87]	0.370 [0.33,0.41]	0.277 [0.24,0.32]	0.000 [0.00,0.00]	0.047 [0.03,0.07]
depth 4	65	0.831 [0.74,0.92]	0.636 [0.57,0.70]	0.547 [0.48,0.61]	0.026 [0.01,0.05]	0.078 [0.05,0.11]
held lexicon	500	0.870 [0.84,0.90]	0.607 [0.58,0.63]	0.512 [0.48,0.54]	0.022 [0.02,0.03]	0.073 [0.06,0.09]
surface	74	0.429 [0.36,0.51]	0.433 [0.39,0.47]	0.350 [0.31,0.39]	0.018 [0.01,0.02]	0.042 [0.03,0.05]
closure	269	0.888 [0.86,0.92]	0.694 [0.67,0.72]	0.591 [0.56,0.62]	0.016 [0.01,0.02]	0.050 [0.04,0.06]

Table 20: Full ordered-pair composition matrices per organism (post-hoc descriptive view of the validated development-suite rows). Row = first operator applied, column = second. Each cell is the two-stage CIIA_2 over the frozen evaluation-plan rows carrying that signature (rows and included seeds averaged within base, then bases uniformly), with the number of bases in parentheses; signatures live in `test_pairs_iid`, the held-signature `test_pair` pool (†), or the order-holdout `test_order` pool (‡). "—" marks a cell with no eligible frozen row: same-operator words are involutive (target equals base, undefined) and no valid NUM/PASS or PASS/NUM depth-2 word exists in the released test pools; such cells are masked, never zeroed. The pair-holdout and order-holdout cells are the only held signatures inside the active set.

Organism	first \ second	NUM	TENSE	POL	BIND	PASS
positive control ($n_s = 1$)	NUM	—	1.00 (53)	0.83 (36)	1.00 (28)	—
	TENSE	1.00‡ (104)	—	0.91 (86)	1.00 (45)	0.78 (9)
	POL	0.82 (40)	0.73 (78)	—	0.74† (92)	0.09† (43)
	BIND	1.00 (31)	1.00 (44)	0.91† (89)	0.15 (20)	1.00 (6)
	PASS	—	0.44 (9)	0.41† (41)	0.43 (7)	—
standard ($n_s = 5$)	NUM	—	0.74 (53)	0.77 (36)	0.94 (28)	—
	TENSE	0.67‡ (104)	—	0.60 (86)	0.79 (45)	0.69 (9)
	POL	0.42 (40)	0.37 (78)	—	0.59† (92)	0.38† (43)
	BIND	0.86 (31)	0.75 (44)	0.85† (89)	0.95 (20)	0.93 (6)
	PASS	—	0.38 (9)	0.62† (41)	0.26 (7)	—
shortcut ($n_s = 5$)	NUM	—	0.66 (53)	0.61 (36)	0.69 (28)	—
	TENSE	0.47‡ (104)	—	0.54 (86)	0.74 (45)	0.53 (9)
	POL	0.42 (40)	0.35 (78)	—	0.41† (92)	0.23† (43)
	BIND	0.70 (31)	0.69 (44)	0.73† (89)	0.84 (20)	0.97 (6)
	PASS	—	0.20 (9)	0.55† (41)	0.23 (7)	—
shuffled ($n_s = 5$)	NUM	—	0.01 (53)	0.00 (36)	0.00 (28)	—
	TENSE	0.00‡ (104)	—	0.00 (86)	0.00 (45)	0.00 (9)
	POL	0.01 (40)	0.00 (78)	—	0.00† (92)	0.02† (43)
	BIND	0.00 (31)	0.00 (44)	0.00† (89)	0.00 (20)	0.00 (6)
	PASS	—	0.04 (9)	0.00† (41)	0.00 (7)	—
random ($n_s = 5$)	NUM	—	0.06 (53)	0.04 (36)	0.06 (28)	—
	TENSE	0.06‡ (104)	—	0.06 (86)	0.04 (45)	0.02 (9)
	POL	0.06 (40)	0.04 (78)	—	0.05† (92)	0.01† (43)
	BIND	0.10 (31)	0.05 (44)	0.07† (89)	0.02 (20)	0.03 (6)
	PASS	—	0.00 (9)	0.00† (41)	0.00 (7)	—

organism the matrix is far from uniform: words that apply BIND or NUM first stay closest to the unary level, whereas words that apply POL or PASS first fall well below it. Table 14's pooled CIIA_2 therefore averages over strongly heterogeneous signatures, and the first operator applied matters more than the second. Seed-level cells vary widely (the artifact records them per seed).

Table 21: Order-holdout path-dependence analysis (post-hoc descriptive). The held order (TENSE,NUM) in `test_order` is compared with the seen order (NUM,TENSE) in `test_pairs_iid` on the bases that carry both signatures in the frozen plan. The two toggles commute, so both orders share one target continuation; any difference is neural path dependence of the composed frozen operators. Δ = held $-$ seen with a base-level paired percentile bootstrap (10,000 replicates, seeds averaged within base) and a sign-flip p; the last column is the within-seed rate at which the two orders disagree in correctness on a shared base.

Organism	shared bases	held CIIA	seen CIIA	Δ [95% CI]	p	disagreement
positive control	28	1.000	1.000	$+0.000$ $[+0.000, +0.000]$	$p = 1.000$	0.000
standard	28	0.693	0.721	-0.029 $[-0.079, +0.014]$	$p = 0.385$	0.086
shortcut	28	0.529	0.664	-0.136 $[-0.200, -0.071]$	$p < .001$	0.150
shuffled	28	0.000	0.000	$+0.000$ $[+0.000, +0.000]$	$p = 1.000$	0.000
random	28	0.043	0.043	$+0.000$ $[+0.000, +0.000]$	$p = 1.000$	0.000

Table 22: Explanation classes on the replication organisms (exploratory). Each class is a set of five operator maps: every operator at one uniform rank at the frozen site, the frozen per-operator selection, or the selection on the SHORTCUT organism. Columns are five-seed means of base-uniform accuracies: unary U, closure-local CL (the second operator against its image on the reached state), closure endpoint CE, held-signature composition H (held pairs, triples and depth-4 words) and IID composition I; *gate* is the candidate-mass gate on the unary and held pools on every seed; $\in M$ marks the unary-matched set (within 0.05 of the best unary accuracy, gate passing) on which the tests run.

class	U	CL	CE	H	I	gate	$\in M$
rank 1	0.446	0.318	0.195	0.069	0.181	pass	
rank 2	0.631	0.437	0.319	0.206	0.295	pass	
rank 4	0.765	0.551	0.450	0.326	0.464	pass	•
rank 8	0.792	0.577	0.486	0.352	0.499	pass	•
rank 16	0.789	0.584	0.488	0.352	0.502	pass	•
rank 32	0.768	0.583	0.497	0.367	0.512	pass	•
selected ranks	0.753	0.534	0.447	0.286	0.458	pass	•
SHORTCUT	0.653	0.503	0.393	0.246	0.391	pass	

Accuracy ranges across the five unary-matched classes: U 0.039 [0.030, 0.053], CL 0.050 [0.039, 0.069], and H 0.081 [0.059, 0.106].

Order holdout. Table 21 pairs the held order (TENSE,NUM) with the seen order (NUM,TENSE) on the 28 bases that carry both signatures. For the standard organism the held order scores 0.693 against 0.721 ($\Delta = -0.029$, 95% CI [-0.079,+0.014], $p = 0.385$), with a within-seed disagreement rate of 0.086; the shortcut organism shows a larger held-minus-seen gap ($\Delta = -0.136$, CI [-0.200,-0.071], $p < .001$). Because the two toggles commute, both orders share one target continuation, so these differences measure neural path dependence of the composed frozen operators, not a linguistic order effect.

H.2 Variation across explanation classes

Explanation classes with matched unary accuracy. Table 22 compares closure-local and held-signature accuracy across explanation classes with similar unary accuracy. The classes are every operator refitted at one uniform rank at the frozen site, the frozen per-operator selection, and the selection on the SHORTCUT organism, all on the replication organisms. The unary-matched, gate-passing set holds five classes (rank 4, rank 8, rank 16, rank 32, selected ranks). Across them unary accuracy ranges over 0.039, closure-local accuracy over 0.050 and held-signature composition over 0.081.

I Natural-language design, results, and annotation

I.1 Completed Family-A design

The completed experiment tests frozen TENSE and NUM interchanges on controlled researcher-generated English stimuli. Source–base pairs share a lexical frame and differ in exactly the target

variable. A tense adverb or subject number supplies the donor information before the intervention site. The tokenizer-specific pre-site-cue check verifies that donor and base prefixes differ for both operators on all 128 target items. Candidate continuations are truncated to the shortest common token length at which they are pairwise distinct. These equal-length minimal distinguishing prefixes separate passive forms at the auxiliary and active-present forms at the verb, allowing the readout to distinguish the prescribed edits.

Both orders of depth-2 TENSE×NUM composition are evaluated on two lexical frames held out from fitting and selection. The high-level toggles commute in this fragment, so the two orders test whether the frozen maps preserve the intended endpoint along both paths. Each applied operator must change the scored candidate. Active-past NUM is syncretic, as in *served* for both numbers, and is undefined on those 32 target items and on every composed row whose number step lands there. Eligible rows are scored under the final Family-A inclusion mask from round-3 human annotation (Appendix I.4.2).

Layer, position, and rank are selected on validation pools. For each model, operator-fitting seed 0 selects a common layer, nominal span, and operator ranks; seeds 1 and 2 reuse that selection with model weights fixed. Table 23 reports the selected configurations, capability checks, and test results. The stimulus and annotation details below document the evaluation items, and Appendix I.6 explains the choice of the number–tense pair.

I.2 Model results and closure-local behavior

I.2.1 Family-A model results

Table 23 reports every validated run under the number–tense protocol and its model-panel extension, which adds three models with the remaining protocol fixed. The panel comprises Pythia-1.4B, Pythia-2.8B, Pythia-410M, Qwen2.5-1.5B, SmolLM2-1.7B, SmolLM2-360M, SmolLM3-3B, spanning three families and 360M to 3B parameters, with 14.3 GPU-hours of run time on an NVIDIA GeForce RTX 4090. The runs use the tokenizer-specific site checks and frozen selections described in Appendix I.1. The table reports two-stage CIIA on the held-out test frames, using equal-length candidate readout and excluding undefined active-past number rows from n.

The results have status **final (round-3 inclusion mask applied)**. The round-3 Family-A inclusion mask retains every Family-A target item, so applying it removes no row from the evaluated pools. The round-3 adjudication supplies the final inclusion decisions. Appendix I.4.2 gives the agreement results and the Family-B instruction deviation, together with the unchanged Family-A instructions and inclusion rule.

Six models attain mean composed CIIA of at least 0.889 in both orders. SmolLM2-360M and SmolLM3-3B sit at ceiling on every evaluation pool of the number–tense protocol. Pythia-1.4B, SmolLM2-1.7B, and Qwen2.5-1.5B have composed means between 0.889 and 0.917. For Pythia-2.8B, validation selects layer 0. Its rank-8 NUM map does not converge on one of three fitting seeds (fit loss 7.1 against 0.13 for the other two, which are at ceiling). All three seeds remain in the reported means, giving 0.667 for TENSE followed by NUM, 0.639 for the reverse order, and 0.667 for unary NUM. All seven models pass the capability and candidate-mass gates. The table retains the unary results, seed variability, and item-level intervals alongside the composition results.

A rank-bounded interchange at one site realizes both operators and their compositions across three families and sizes from 360M to 3B, with differences between models within the item-level intervals over 24 to 32 items per pool.

I.2.2 Closure-local decomposition on the pretrained models

The closure-local analysis reconstructs the intermediate outcomes needed to score the second operator locally. The original run artifacts contain endpoint correctness. Each run's two unary maps are refitted at its frozen layer, span, and ranks with the original fitting configuration and seed. Recorded fit losses and every frozen unary and endpoint row serve as reproduction checks. Each composed row is then scored without intervention, after the first operator, and after both. The local target is the second operator's high-level image on the candidate actually reached by the first. Table 24 com-

Table 23: Family-A natural-language interchange results under the number–tense protocol (TENSE×NUM, depth 2, both orders; equal-length candidate scoring; test frames only). Two-stage CIIA, mean±sd over operator-fitting seeds, with a 95% percentile bootstrap interval over base items of the seed-averaged item scores; n is the number of defined rows (syncretic active-past number rows are undefined, never forced); behav. is the smallest across seeds of the base model's observational accuracy on the same candidates (behavioral-capability gate, threshold 0.8); $P(\mathcal{C})$ is the smallest across seeds of the intervened-to-observational candidate-mass ratio (candidate-mass gate, threshold 0.5). Pool names abbreviate the two unary operators and the two composition orders. Site: layer/span; ranks: NUM/TENSE; three fitting seeds per model. Pools contain the Family-A items retained by the round-3 human inclusion mask. † added by the panel extension under the same protocol.

model	site	ranks	pool	n	CIIA [95% CI]	behav.	$P(\mathcal{C})$
Pythia-1.4B	L14/m1	N1/T2	TENSE	32	0.969 [0.91,1.00]	0.969	1.05
			NUM	24	1.000 [1.00,1.00]	1.000	1.01
			T→N	24	0.917 [0.79,1.00]	0.958	1.06
			N→T	24	0.917 [0.79,1.00]	0.958	1.08
Pythia-2.8B†	L0/m4	N8/T1	TENSE	32	1.000 [1.00,1.00]	1.000	1.30
			NUM	24	0.667±0.471 [0.67,0.67]	1.000	0.99
			T→N	24	0.667±0.471 [0.67,0.67]	1.000	1.29
			N→T	24	0.639±0.453 [0.60,0.67]	1.000	1.29
Pythia-410M	L18/m1	N8/T2	TENSE	32	1.000 [1.00,1.00]	1.000	1.06
			NUM	24	0.944±0.039 [0.86,1.00]	1.000	1.00
			T→N	24	1.000 [1.00,1.00]	1.000	1.06
			N→T	24	1.000 [1.00,1.00]	1.000	1.05
Qwen2.5-1.5B†	L22/m1	N1/T1	TENSE	32	0.917±0.059 [0.83,0.98]	1.000	1.00
			NUM	24	1.000 [1.00,1.00]	1.000	1.01
			T→N	24	0.903±0.071 [0.81,0.97]	1.000	0.99
			N→T	24	0.889±0.079 [0.78,0.97]	1.000	0.99
SmolLM2-1.7B	L18/m1	N2/T1	TENSE	32	0.875 [0.75,0.97]	0.938	0.98
			NUM	24	1.000 [1.00,1.00]	0.917	0.97
			T→N	24	0.917 [0.83,0.99]	0.917	0.95
			N→T	24	0.917 [0.83,0.99]	0.917	0.95
SmolLM2-360M	L12/m4	N2/T1	TENSE	32	1.000 [1.00,1.00]	1.000	0.94
			NUM	24	1.000 [1.00,1.00]	1.000	1.16
			T→N	24	1.000 [1.00,1.00]	1.000	1.14
			N→T	24	1.000 [1.00,1.00]	1.000	1.17
SmolLM3-3B†	L21/m1	N1/T2	TENSE	32	1.000 [1.00,1.00]	1.000	0.79
			NUM	24	1.000 [1.00,1.00]	1.000	1.24
			T→N	24	1.000 [1.00,1.00]	1.000	1.03
			N→T	24	1.000 [1.00,1.00]	1.000	1.02

pares this local accuracy with unary accuracy on the same measurable items, following the synthetic decomposition in Table 11.

This descriptive analysis uses one commuting pair at depth 2 and the final Family-A inclusion mask from round-3 human annotation. The composed pools contain 24 items per order, and the local comparisons retain 8 to 9 measurable items per model and second operator. Reproduction checks pass for 21 of 21 refits, covering the recorded losses, unary rows, and endpoint rows.

I.3 COPULA AGREEMENT TEST ON THE PRETRAINED MODELS

On passive frames the copula's form depends on both number and tense (*is*, *are*, *was*, *were* at the first scored token). A second step that reads the subject's number from its token cue, rather than from the intervened subspace, therefore decodes a stale form. The copula agreement test scores the number–tense pair on this contrast at scale. Its materials use a closed lexicon screened for hygiene, with 305 test nouns and their own verbs and agents, disjoint from the frames on which the maps were

Table 24: Closure-local decomposition of the natural-language composed pools (descriptive). For each model and second operator: the second operator's accuracy scored against its own image on the candidate the first operator actually reached (CL), the same operator's unary accuracy on the same items, and their paired difference, each a mean over three fitting seeds, with the number n of distinct items: the composed pool's items that also carry a unary row for that operator, minus rows whose reached candidate is shared by two cells (the syncretic active past), on which the second operator's image is not a single candidate; positive and negative passive twins share the scored prefix and receive identical rows, so each pair counts once. *pooled* averages each item over the orders in which it is measurable. *repr.* counts fitting seeds whose refit losses, unary rows and endpoint rows all reproduced the frozen run artifact. Pools hold eight or nine distinct items, so the table reports point estimates. The analysis uses one commuting pair at depth 2 and the final Family-A inclusion mask from round-3 human annotation.

model (repr.)	second operator	n	closure-local	unary	gap
Pythia-410M (3/3)	NUM	8	1.000	0.917	+0.083
	TENSE	9	0.815	1.000	-0.185
	pooled	9	0.852	0.963	-0.111
Pythia-1.4B (3/3)	NUM	8	0.875	1.000	-0.125
	TENSE	9	0.778	1.000	-0.222
	pooled	9	0.778	1.000	-0.222
Pythia-2.8B (3/3)	NUM	8	0.667	0.667	+0.000
	TENSE	9	0.889	1.000	-0.111
	pooled	9	0.722	0.852	-0.130
SmolLM2-360M (3/3)	NUM	8	1.000	1.000	+0.000
	TENSE	8	1.000	1.000	+0.000
	pooled	8	1.000	1.000	+0.000
SmolLM2-1.7B (3/3)	NUM	8	0.875	1.000	-0.125
	TENSE	8	0.875	0.750	+0.125
	pooled	8	0.875	0.875	+0.000
SmolLM3-3B (3/3)	NUM	8	1.000	1.000	+0.000
	TENSE	8	1.000	1.000	+0.000
	pooled	8	1.000	1.000	+0.000
Qwen2.5-1.5B (3/3)	NUM	8	0.958	1.000	-0.042
	TENSE	8	0.833	0.833	+0.000
	pooled	8	0.896	0.917	-0.021

fitted. Each noun contributes the four passive positive cells, giving 1,220 items per word; the 576 items whose subject noun is a single token in both numbers under every tokenizer form the primary stratum. The materials, intervention words (T, N, TN, NT, TNT, NTN), estimators, and decision rule were specified before evaluation.

Each model's frozen NUM and TENSE maps are refitted from its own run artifacts at its frozen layer, span and ranks with the recorded seeds. A model is scored only if the refit reproduces the recorded fit losses within 10^{-3} and every row of the earlier closure-local decomposition. Every step is scored against its own image on the candidate the previous step actually reached. The contrasts are the item-mean closure-local minus unary difference of the last operator of NT, TN, TNT and NTN, with item and noun-cluster bootstraps, a one-sided cluster sign-flip p and Holm over the four contrasts.

Six models pass the reproduction gate; SmolLM3-3B's refit differs by 0.011 on one seed and is not scored. The loss is confirmed in five of the six (Table 25). Per model it reaches 0.127 (Pythia-1.4B, NUM after TENSE), and SmolLM2-360M stays at ceiling. Pooled over the six models, a descriptive summary because the declared pool required all seven, Δ is -0.047 [-0.057, -0.038] for TENSE after NUM and -0.038 [-0.046, -0.031] for NUM after TENSE, with -0.039 and -0.034 at depth three. Of 109,636 scored steps whose previous step decoded correctly, 8,908 leave the candidate unchanged and 1,872 decode the stale form of the first variable (Table 26).

Pythia-6.9B was added as a scale extension using a validation grid of layers 6, 12, 19 and 25, spans 1 and 4, and ranks 1 to 8; it selects layer 19, span 1, and rank 2 for both operators. Its base-run pools are all at 1.000. On the copula test it gives $+0.004$ [-0.003, $+0.012$] for TENSE after NUM, no loss at ceiling, and -0.008 [-0.014, -0.002] for NUM after TENSE; 70 of its 21,816 steps are stale.

Table 25: Copula agreement test on the pretrained models (primary stratum: subject nouns single-token in both numbers under every tokenizer; 576 items common to all models). For each model and each word ending in the operator shown, the second (or third) operator's closure-local accuracy CL against its own image on the state the previous step reached, the same operator's unary accuracy U on the same items, and the item-mean difference $\Delta = \mathrm{CL} - U$ with its item-bootstrap 95% interval; seed means over three fitting seeds. Pooled row: item means over the models, Holm over the four contrasts, one-sided cluster sign-flip p. Labels follow the declared decision rule; a model whose unary accuracy is at ceiling cannot show a gain. Prospective-exploratory: items are automatically screened and unannotated. Decisions are per model under the declared rule; the pooled row is over the 6 models that passed the reproduction gate (SmolLM3-3B did not reproduce the earlier decomposition within tolerance on one seed and is not scored), computed with the declared estimators.

model	TENSE after NUM			NUM after TENSE			TENSE after NUM, TENSE			NUM after TENSE, NUM		
	U	CL	Δ [95% CI]	U	CL	Δ [95% CI]	U	CL	Δ [95% CI]	U	CL	Δ [95% CI]
Pythia-410M	1.000	0.941	-0.059 [-0.073, -0.046]	0.941	1.000	+0.059 [+0.046, +0.073]	1.000	0.932	-0.068 [-0.085, -0.052]	0.941	1.000	+0.059 [+0.046, +0.073]
SmolLM2-360M	1.000	0.998	-0.002 [-0.005, +0.000]	0.998	0.999	+0.001 [-0.002, +0.005]	1.000	0.996	-0.004 [-0.009, -0.001]	0.998	1.000	+0.002 [+0.000, +0.005]
Pythia-1.4B	0.863	0.818	-0.045 [-0.074, -0.014]	0.999	0.873	-0.127 [-0.149, -0.105]	0.863	0.849	-0.014 [-0.039, +0.011]	0.999	0.897	-0.102 [-0.122, -0.084]
SmolLM2-1.7B	0.947	0.878	-0.069 [-0.092, -0.046]	0.986	0.904	-0.082 [-0.102, -0.061]	0.947	0.895	-0.052 [-0.074, -0.030]	0.986	0.899	-0.087 [-0.106, -0.067]
Qwen2.5-1.5B	0.922	0.821	-0.101 [-0.127, -0.076]	0.994	0.911	-0.083 [-0.101, -0.066]	0.922	0.835	-0.087 [-0.108, -0.065]	0.994	0.914	-0.080 [-0.097, -0.065]
Pythia-2.8B	1.000	0.992	-0.008 [-0.012, -0.004]	0.632	0.634	+0.002 [-0.003, +0.007]	1.000	0.988	-0.012 [-0.018, -0.007]	0.632	0.635	+0.003 [-0.001, +0.007]
pooled (6 models)	0.955	0.908	-0.047 [-0.057, -0.038]	0.925	0.887	-0.038 [-0.046, -0.031]	0.955	0.916	-0.039 [-0.047, -0.032]	0.925	0.891	-0.034 [-0.041, -0.028]
reading	loss, $p_{\mathrm{Holm}} < 0.001$			loss, $p_{\mathrm{Holm}} < 0.001$			loss, $p_{\mathrm{Holm}} < 0.001$			loss, $p_{\mathrm{Holm}} < 0.001$		
Pythia-6.9B (scale extension)	0.992	0.996	+0.004 [-0.003, +0.012]	1.000	0.992	-0.008 [-0.014, -0.002]	0.992	0.996	+0.004 [-0.002, +0.011]	1.000	0.987	-0.013 [-0.021, -0.006]
reading	no loss at ceiling			loss			no loss at ceiling			loss		

Table 26: Failure modes of the closure-local step in the copula agreement test. Rows are (item, word, seed) steps $k \geq 2$ of the words TN, NT, TNT, NTN whose previous step decoded correctly and whose closure-local target is well defined; a failing step leaves the decoded candidate unchanged (no effect), decodes the form that carries the first variable's old value (stale), or returns to the base; distinct (item, word) keys in parentheses. Descriptive; no decision rule.

model	rows	scored	correct	no effect	stale	base	other
pythia-1.4b	14640	19282	17073	1543 (727)	666 (416)	0	0
pythia-2.8b	14640	9832	6967	2825 (1888)	18 (18)	22 (22)	0
pythia-410m	14640	21520	21275	241 (126)	4 (2)	0	0
qwen2.5-1.5b	14640	20256	18225	1489 (850)	539 (378)	3 (3)	0
smollm2-1.7b	14640	20825	19430	840 (527)	553 (369)	2 (2)	0
smollm2-360m	14640	17921	15859	1970 (685)	92 (42)	0	0
pooled	87840	109636	98829	8908	1872	27	0
pythia-6.9b (scale extension)	14640	21816	21635	110 (82)	70 (43)	0	1 (1)

I.4 STIMULUS INVENTORY AND ANNOTATION

I.4.1 FAMILY-A CANDIDATE DESIGN AND INCLUSION

Candidate minimal sets are built around the finite-verb *region*, not a single-token three-variable grid. In active simple past, singular and plural are syncretic both affirmatively (*ate*) and under past do-support (*did not eat*); NUM is therefore undefined in those cells and cannot enter a pair metric. POL is a span operation (e.g., *eats* ↔ *does not eat*), not a one-token label. In passives, verbal number agrees with the surface subject/theme rather than the postverbal agent. Identifiable cells otherwise use agreement, do-support, and auxiliary selection: *the chef eats / the chefs eat / the chef does not eat / the chefs do not eat / the chef ate / the chefs did not eat* (and passive counterparts with *is/are/was/were*). The released candidate packet is exactly balanced over eight lexical frames, voice, singular/plural, tense, and polarity. The candidate design targeted token-length, animacy, and lemma-frequency matching. The diagnostic *span* must be shown absent from each model prefix by tokenizer-specific leakage checks. Sixteen unmarked single-error controls cover agreement, do-support, double-past, and passive auxiliary/participle errors for behavioral capability checks, not interventions. There is no admissible all-grid CIIA_2 endpoint.

The round-1 inventory contains 128 balanced Family-A factorial targets, 16 single-error controls, and 32 Family-B items. Rounds 2 and 3 retain the Family-A material, re-author the Family-B entail-

ment items, and add 16 planted-ambiguity controls. Agreement is computed from the independent A/B ratings before adjudication. A Family-A item enters the inclusion mask when A/B agreement meets the stated thresholds and adjudication finds no multiple reading; error subtype, confidence, and comments are retained as diagnostics.

I.4.2 ROUND-3 OUTCOME

Three annotators new to the study, two scoring and one adjudicating, returned the round-3 workbooks on 2026-09-02. The final annotation round met the stated agreement criteria on all ten dimensions: in Family A, quadratic-weighted $\kappa = .872$ for graded acceptability and exact agreement 1.000 for surface number, tense, polarity and voice over 144 items; in Family B, $\kappa_w = .828$ and $.934$ for naturalness and plausibility and exact agreement 1.000 for the three premise–hypothesis relations over 64 items. The planted-ambiguity instrument passed: each scorer flagged 16 of the 16 planted items against a threshold of 13, with Gwet's $\text{AC}_1 = 1.000$ against a threshold of $.80$.

The Family-B naturalness instructions departed from the frozen packet: the scale's definition and anchors were clarified to all three annotators before they rated (naturalness rates the English string and construction, not the plausibility of the described event); every Family-A codebook row, the relation rows, and all stimuli, hypotheses and per-role orders are identical to the frozen packet. Family-B naturalness, plausibility and the item licensing that consults them are therefore treated as exploratory and no Family-B transfer number is reported, while Family A keeps its computed status. Round 3 was the final annotation round.

Annotators and determination. The three annotators are project collaborators credited as co-authors and received no monetary payment; their consent and roles are recorded in the study's human-subjects readiness document. The responsible investigator made a written determination, dated 2026-07-25, that the work is professional linguistic annotation by collaborators rather than human-subjects research.

I.5 FAMILY-B CAUSED-MOTION DESIGN

Family B provides caused-motion stimuli and full premise–hypothesis annotations. Naturalness, plausibility, and item licensing are reported as exploratory under the instruction deviation in Appendix I.4.2. No Family-B transfer number is reported.

The design contrasts typical motion-compatible verbs (e.g., *push, throw, kick, pull*) with manner verbs receiving a caused-motion construal in the construction (e.g., *laugh, sneeze, talk, wave*). The verb roots are not claimed to entail caused motion by themselves. Examples include *They laughed him out of the room* / *She sneezed the foam off the cappuccino*. Entailment is elicited directly for each complete premise–hypothesis pair, including affected-theme displacement and path endpoint hypotheses, with an uncertainty option and later adjudication. For each probe dimension, condition, and verb class, two items use a positive hypothesis and two its cancellation-form probe. All three annotators see the same premise and three hypotheses for an item; only item order differs. Continuation preferences (e.g., *...and he was no longer in the room* versus a manner-only continuation) are a separate graded preference diagnostic; they are never converted into entailment ground truth. Construction-absent and implausibility controls are included. Eight target verb–path exposure probes are not novelty or training-data-absence controls. In the frozen UD 2.18 screen, the form pairs *push–off* and *kick–out of* are observed; the other six are only unobserved within that 1.11M-word corpus/query and require a larger-corpus sensitivity check. No motion-entailment label is admitted before human annotation of the full pair.

Capability controls identify the caused-motion frames supported by the base model and thus eligible for testing their neural implementation.

I.6 PROTOCOL DESIGN

Annotation ran in three rounds, and the round-3 inclusion mask, which passed every Family-A gate on 2026-09-02, determines the reported model results. Appendix I.4 records the inventory and Appendix I.4.2 the agreement scores.

Table 27: Surface-alignment sensitivity on matched standard-model rows.

	parent	perturbed, right-aligned	perturbed, role-aligned
CIIA	0.724	0.435	0.684
agreement with parent	–	0.643	0.866

The number–tense pair is the pair the prefix-site protocol can test. English do-support negation supplies no cue before the finite verb, so for a polarity interchange the base and donor prefixes are token-identical and copying donor activations leaves the base unchanged; the tense adverb and the subject number, by contrast, make the donor prefix differ from the base on all 128 target items, and the pre-site-cue gate rejects any operator whose donor and base prefixes are identical. Scoring uses equal-length minimal distinguishing prefixes, so candidates of different length are never compared. Active-past NUM is undefined under the rule that each step must change the scored candidate, as specified in Appendix I.1.

J Additional protocol and representation analyses

These descriptive development-suite analyses examine surface alignment, subspace stability, and map complexity using the reproduced models described in Appendix E.2.

J.1 Surface alignment

Surface alignment sensitivity. A target-preserving adverb perturbation lowers standard-model CIIA from 0.724 to 0.435 and likewise lowers the modular reference from 0.877 to 0.431. Aligning donor positions to corresponding non-adverb tokens recovers most of the standard model's loss, raising CIIA to 0.684; the remaining realigned-minus-parent gap is -0.039 (95% CI [-0.067,-0.012]). This identifies token alignment as an important property of the intervention interface while leaving a smaller residual.

J.2 Subspace stability

Table 28 compares repeated operator fits inside the same frozen model before comparing subspaces across model seeds. Within a model, the mean $\cos^2$ of the principal angles to the frozen fit ranges from 0.35 to 0.85 across operators, against a random baseline of at most 0.06. The refits retain nearly the frozen validation CIIA. The learned intervention function is stable within a model, while the rank-r subspace that carries it is not unique.

Across models, the comparison uses raw residual coordinates without a shared basis. The mean $\cos^2$ of the principal angles reaches at most 0.07. Transplanting a subspace into another seed's model gives validation CIIA at most 0.00, compared with at least 0.79 in the model where it was fitted. These results locate the stability in the fitted intervention's behavior within a model, while its subspace coordinates remain model-specific.

J.3 Rank budgets and map families

J.3.1 Rank sweep

Table 29 rescores the frozen evaluation rows with a common rank for every active operator. CIIA_3 rises from 0.382 at rank 1 to 0.717 at rank 16, while closure CIIA changes less across the upper ranks. The validation-selected configuration is shown for comparison.

J.3.2 Map-family comparison

This analysis compares rank-bounded, full-linear, MLP, and null-preserving MLP intervention maps fitted to the same unary-only rows at the frozen site, with optimizer selection on validation data. On the standard organism, the rank-bounded family has composed CIIA 0.689 averaged over IID pairs, IID triples, and closure. The higher-capacity families attain high unary accuracy but fail the

Table 28: Stability of the learned operator subspaces (post-hoc descriptive; standard organism, seeds 0,1,2,3,4). Refit: the same frozen model with the operator refitted under fitting seeds 100,101 (different row subsample and initialization) at the frozen site/rank; mean $\cos^2$ of the principal angles to the frozen fit, averaged over models and refits ($\pm$ sd). Cross-model: frozen subspaces of two different training seeds compared in raw residual coordinates; different seeds share no basis, so this column is expected at the random baseline (mean $\cos^2$ of two independent random subspaces of the same rank in $\mathbb{R}^{128}$). Val CIIA: unary validation CIIA of the frozen fit and of the refits (functional stability). Transplant: the subspace fitted on seed t applied inside seed s's model on the first 40 validation rows; diagonal ($s{=}t$) versus off-diagonal ($s{\neq}t$) means.

		mean $\cos^2$ principal angles			val CIIA		transplant val CIIA	
Operator	rank	refit	cross-model	random	frozen	refits	$s{=}t$	$s{\neq}t$
NUM	1	0.850 ± 0.215	0.008 ± 0.012	0.007	0.920	0.935	0.89 ± 0.14	0.00 ± 0.00
TENSE	8	0.426 ± 0.122	0.070 ± 0.005	0.062	0.792	0.787	0.79 ± 0.26	0.00 ± 0.01
POL	8	0.350 ± 0.119	0.062 ± 0.008	0.062	0.890	0.865	0.88 ± 0.25	0.00 ± 0.00
BIND	8	0.463 ± 0.068	0.060 ± 0.007	0.063	0.960	0.960	0.95 ± 0.11	0.00 ± 0.00
PASS	8	0.409 ± 0.064	0.065 ± 0.010	0.062	0.997	1.000	1.00 ± 0.00	0.00 ± 0.00

Table 29: Exploratory test-set rank sweep for the faithfulness–complexity Pareto view (standard organism, training seeds 0,1,2,3,4, frozen common site layer 0 / nominal span 12, frozen evaluation-plan rows). Every active operator's rank is overridden to one common value; the validation-selected configuration (NUM 1, TENSE/POL/BIND/PASS 8; total rank 33) is the only confirmatory point and is repeated in the last row from the primary suite restricted to the same training seeds (the all-seed primary means are in Table 14). Cells are two-stage CIIA over bases (seeds averaged within base); complexity is total rank $\times$ nominal span. The non-selected ranks form an exploratory test-set response curve.

rank/op	$\sum r_o$	cplx	CIIA_1	CIIA_2	held pair	order	CIIA_3	held triple	CIIA_4	lex	closure	$P(\mathcal{C})@1$
1	5	60	0.510	0.262	0.176	0.487	0.382	0.073	0.311	0.343	0.319	0.950
2	10	120	0.777	0.462	0.405	0.596	0.533	0.303	0.502	0.530	0.538	0.929
4	20	240	0.873	0.589	0.595	0.656	0.644	0.326	0.586	0.606	0.652	0.940
8	40	480	0.902	0.665	0.639	0.700	0.684	0.386	0.640	0.662	0.706	0.950
16	80	960	0.928	0.651	0.649	0.687	0.717	0.363	0.631	0.665	0.708	0.939
32	160	1920	0.912	0.661	0.620	0.673	0.713	0.412	0.656	0.694	0.691	0.930
frozen	33	396	0.888	0.644	0.639	0.669	0.681	0.370	0.636	0.607	0.694	–

candidate-mass gate on at least one composed pool; the null-preserving MLP retains the identity when base and donor are equal.

A direct arm fits rank-bounded maps on depth-2 validation rows and reaches composed CIIA 0.725. A linear probe at the same site decodes the operator fields with accuracy 0.81–1.00 against a majority baseline at most 0.53, indicating that simple decodability is not the limiting factor.

Table 30: Map-family comparison under a prospective protocol, evaluated as an exploratory analysis after the development-suite freeze. Every family is fitted on the same unary-only training rows at the frozen common site with an optimizer selected on validation rows only, then scored on the frozen evaluation-plan rows; the direct arm is the rank-bounded family trained jointly on depth-2 validation rows. Cells are two-stage CIIA (seeds averaged within base). $P(\mathcal{C})$ ratio is the candidate-mass ratio on IID pairs and "gate" says whether every seed passed it; "null" is the mean relative displacement $\|\Phi(h,h) - h\|/\|h\|$ on validation activations (exactly 0 for null-preserving families). The verdict under the map-family rules is **supported**.

map family	organism	seeds	CIIA_1	CIIA_2	CIIA_3	closure	$P(\mathcal{C})$ ratio	gate	null
rank-bounded (primary)	standard	0,1,2,3,4	0.903	0.658	0.690	0.719	0.89	yes	0.00
rank-bounded (primary)	random	0,1,2,3,4	0.053	0.048	0.050	0.050	1.01	yes	0.00
rank-bounded (primary)	shuffled	0,1,2,3,4	0.003	0.002	0.135	0.009	0.95	yes	0.00
full linear	standard	0,1,2,3,4	0.988	0.566	0.305	0.579	0.45	no	0.00
full linear	random	0,1,2,3,4	0.112	0.084	0.052	0.077	1.25	yes	0.00
full linear	shuffled	0,1,2,3,4	0.035	0.028	0.099	0.046	0.58	no	0.00
MLP	standard	0,1,2,3,4	0.988	0.459	0.297	0.548	0.17	no	293.14
MLP	random	0,1,2,3,4	0.229	0.087	0.064	0.081	0.74	no	292.53
MLP	shuffled	0,1,2,3,4	0.064	0.039	0.098	0.032	0.15	no	460.11
MLP, null-preserving	standard	0,1,2,3,4	0.949	0.551	0.358	0.659	0.22	no	0.00
MLP, null-preserving	random	0,1,2,3,4	0.112	0.071	0.045	0.068	1.05	yes	0.00
MLP, null-preserving	shuffled	0,1,2,3,4	0.036	0.038	0.084	0.032	0.37	no	0.00
direct (pairs-trained)	standard	0,1,2,3,4	0.798	0.719	0.767	0.690	0.85	yes	–
direct (pairs-trained)	random	0,1,2,3,4	0.049	0.049	0.049	0.055	0.92	yes	–
direct (pairs-trained)	shuffled	0,1,2,3,4	0.032	0.020	0.165	0.019	0.95	yes	–

K Positioning and study considerations

K.1 Positioning

K.1.1 Causal abstraction and recursive interchange

Causal abstraction and DAS align internal states with high-level variables, and recursive interchange makes interventions on already intervened states formally available (Geiger et al., 2021; 2024; 2025). Type-level IIT induces typed character variables (Huang et al., 2023), while PYVENE supports serial interventions on upstream-intervened runs (Wu et al., 2024). Our advance over these separate precedents is a joint test of rank-bounded linguistic maps fitted independently on unary counterfactuals, then frozen for held-out sequential composition and evaluation on intervention-reached states.

ACDC, EAP/EAP-IG, and causal scrubbing provide circuit-discovery and hypothesis-validation procedures (Conmy et al., 2023; Syed et al., 2024; Chan et al., 2022). The evaluation objects in this paper are independently learned operator implementations, with their mapping class, source policy, and composition domain declared before evaluation.

Independently developed interventions have been applied in sequence and shown to interfere (Kolbeinsson et al., 2025). Multi-attribute interchange benchmarks isolate typed features (Huang et al., 2024), while subnetwork ablation provides causal evidence of structural compositionality (Lepori et al., 2023). With the independently unary-fitted maps frozen, our joint test asks whether each still implements its linguistic operator on states produced by earlier maps.

K.1.2 Explanation complexity and intervention-reached states

Dormant pathways, intervention-induced distribution changes, and unrestricted alignments can affect the interpretation of a successful unary intervention (Makelov et al., 2024; Grant et al., 2025; Sutter et al., 2025; Zhang & Nanda, 2024; Méloux et al., 2026). These issues motivate specifying both the capacity of an explanation map and the states on which it is evaluated. Grant et al. (2025) characterize representation sets closed under patching. Our intervention closure is generated by the declared operator implementations and source rules, and evaluation compares their actions with the high-level operators on those reached states.

Successor heads contain a decodable numeric feature whose edits fail in one direction (Gould et al., 2024). This is a within-head instance of a recovered representation whose edit misses the intended target.

Formal validation connects interpretation to semantic guarantees. Axiomatic work requires a mechanistic interpretation to preserve semantics compositionally (Palumbo et al., 2025), while certified circuit discovery supplies robustness, patching, and minimality guarantees over continuous regions (Hadad et al., 2026). Theorem 1 gives sufficient conditions for propagating local operator error along an admissible intervention process, with the one-step guarantee imposed on the required intervention closure. The geometric factor $\sum_{j=0}^{k-1} L^j$ is attained by the tightness construction in Appendix B.2.

K.1.3 Linguistic operators and compositional computation

Capability accounts characterize what Transformers can compute (Weiss et al., 2021; Lindner et al., 2023; Hahn, 2020; Strobl et al., 2024). Mathematical linguistics supplies typed operations, state updates, constructions, and role–filler binding (Montague, 1970; Lambek, 1958; Heim, 1982; Goldberg, 1995; Smolensky, 1990). We use these established computational and semantic objects to specify the transformations and validity conditions that a neural implementation is evaluated against.

Composition has been studied through circuit-mask union (Mondorf et al., 2025), qualitative DII/PEFT subspace combination (Liao & Monz, 2024), and reusable task interfaces (Ma et al., 2026). The causal MDL frontiers of Ma et al. (2026) test unseen task compositions. These methods compose masks, subspaces, or task interfaces. Our comparison follows sequential applications of separately fitted linguistic maps to the activations produced by earlier maps. Concurrent work provides a supervised typed mechanism library with exact interventional ground truth and legal edit sequences (Xun, 2026). Its designed slots supply explicit mechanisms, with edits that do not control

answer readout. OpClosure supplies linguistic counterfactuals for evaluating learned intervention maps and their effects on the declared readout.

Function-vector arithmetic composes task controls (Todd et al., 2024), while steering studies target tense and aspect (Klerings et al., 2025).

Mechanistic studies identify a causal circuit for function composition (Tang et al., 2026), decode intermediate values in two-hop factual composition (Khandelwal & Pavlick, 2026), and measure representational homomorphism error (An & Du, 2026). Constituent-aware pooling probes where phrasal meaning composes (Aljaafari et al., 2026). Linguistic interventions separately study agreement, binding, filler–gap dependencies, logical predicates, and Dyck hierarchy variables (Finlayson et al., 2021; Feng & Steinhardt, 2024; Arora et al., 2024; Boguraev et al., 2025; Chauhan et al., 2026; Sharma et al., 2026). These studies analyze compositional computation, representations, and individual linguistic phenomena. Our advance is a joint test of typed grammar-state transformations through independently unary-fitted, rank-bounded maps, frozen for held-out sequential composition and evaluation on intervention-reached states.

K.2 Study considerations

K.2.1 Declared implementations and observable outcomes

A realization claim specifies a model, an operator inventory, an explanation class, an intervention site, a source policy, and an observable signature. Theorem 1 supplies sufficient conditions under which closure-level one-step control bounds composition error for that implementation. Site selection uses validation pools to choose among the tested sites. Agreement with the declared signature permits multiple internal implementations. Table 28 reports refits with similar validation behavior and different subspace orientations.

The protocol uses rank as a proxy for explanatory complexity. Proposition 2 states the classical description-length bound, while the implemented budget controls writable directions through rank and span. The candidate-set signature $\mathcal{C}$ determines which outcome differences can be observed. Agreement within $\mathcal{C}$ can coexist with divergence outside it, so the evaluation reports candidate mass alongside the scores on the declared signature.

K.2.2 Interpreting the modular reference

The modular reference is trained with unary interchange targets at the designated layer-1, span-8 oracle site (Appendix D.2). The development-suite comparison uses one reference seed; a supplementary five-seed run shows moderate variation ($\mathrm{CIIA}_3 = 0.764 \pm 0.05$). The replication and further-instance analyses in Appendix G calibrate the closure-local deficit against this reference.

K.2.3 Language fragment and intervention interface

The synthetic grammar uses English-like morphology and one construction alternation. Recursion is limited to a single filler–gap dependency. It contains no scope, coreference, or discourse-state obligations. Themes are constrained to be singular, and validity excludes caused-motion passives and wh-extraction from passives (Appendix C.3). Object WH remains frame-only because questions have no echo tail, while PASS is an active intervention (Appendix D.1). The development-suite order-holdout pair commutes at the outcome level, with 100% agreement on the released data, so the split tests intervention-path dependence. The development-pool depth analysis and the surface-alignment comparisons are reported in Appendices C.3.6 and J.1.

K.2.4 Broader impact

The protocol aims to make interpretability claims explicit about the implementation, evaluation domain, and controls used. A foreseeable misuse is to report that a model passed a composition test while allowing unrestricted alignments or omitting closure states. Coupling the released metric code to its control suite helps preserve the conditions needed to interpret that claim.