WHERE DOES AUGMENTATION REPAIR COMPOSITIONAL GENERALIZATION? THE LOCUS SHIFTS WITH COMPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Augmentation is known to repair compositional generalization failures in neural decoders, but *what* it repairs, and *where*, is less clear. In a controlled toy regime we establish a full mechanism including a sharp repair-locus transition; in more realistic settings we find a weaker but consistent regularity. We study this in a controlled setting: a compressed (superposed) representation is decoded by a serialized language readout, and augmentation is applied under pre-registered, budget-matched protocols (1000+ trained models, all judgments fixed before execution). Four results emerge. (R1) Failure is a content-addressing mismatch. Probes recover the information (separation ratio ≈ 0.96) while slot-level diagnostics locate the collapse in the object slots (0.49–0.74), with the relation slot substantially more robust (0.77–0.96); the pattern holds in both a toy regime and a semi-synthetic bridge. (R2) Repair is unit-local and thresholded. Augmenting a combination rebuilds that combination's readout pathway: 300 examples per unit raise covered units to 0.70–1.00 (median 0.93) while 75 per unit leave accuracy near chance (≈ 0.2), and no transfer to unaugmented units occurs. Coverage, not augmentation content, is therefore the driver of repair: mechanism-guided selection is never better than random in the toy regime, and uniform coverage dominates mechanism-guided selection in the bridge (both pre-registered, 5 seeds). (R3) Compression deepens failure (0.000 at $C{=}4$ versus 0.195 with no compression) and shapes repair efficiency non-monotonically: mid-compression ($\delta \approx 1.5$–2.25) repairs slowest. (R4) In the toy regime the repair locus depends on compression. At $\delta = 2$ the readout pathway alone is sufficient (freezing the encoder still repairs covered units, $\Delta = +0.54$, 5/5 seeds); at $\delta = 4.5$ encoder reorganization is required (freezing costs 0.52, 0/5). A seven-point frozen-encoder scan shows the transition is continuous (crossover between $\delta = 2.6$ and $\delta = 2.2$), and R3's slow-repair plateau disappears when the encoder is frozen: the plateau is a property of encoder reorganization, not of the readout. A first-order account based on Cover's linear-separability capacity predicts the transition band from first principles, with no fitted parameters. On continuous features the locus does not transfer naively (a random frozen code fails by 0.56), while freezing a trained encoder with decoder-only augmentation recovers most of the repair (0.62 vs. 0.75) at lower cost. Under end-to-end training of a small CNN the failure pattern and the weak form both reproduce: unseen accuracy is 0.020 at baseline with object slots at 0.39 against a relation slot at 1.00, and freezing the trained encoder recovers 88.5% of the repair gain (0.539 vs. 0.606). Throughout, geometry changes globally under augmentation while behavior repairs locally: geometric reorganization is incidental at mild compression and is the repair itself at strong compression, a locality that slot-level measurement reveals and geometric analysis cannot.

1 INTRODUCTION

Neural decoders trained on compositional input can fail on combinations of seen components, a phenomenon known as *compositional generalization* (Lake & Baroni, 2018; Keysers et al., 2020; Ruis et al., 2020). A growing line of work shows that augmentation (adding training examples for the

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.

missing combinations) repairs the failure, and that the repair is driven by *coverage*: with sufficient examples per combination, performance recovers (Mahajan et al., 2025; Uselis et al., 2025).

But coverage is a *sufficiency* statement, not a *mechanism*. Two questions remain open. First, what does augmentation repair: the missing joint distribution, or the decoder's readout pathway for a combination? Second, where does it repair: in the representation (encoder geometry) or in the readout (decoder pathway)? These questions matter because they determine whether augmentation can transfer (repair one combination and gain others), whether frozen or full fine-tuning is required, and how repair interacts with representation quality (e.g., superposition).

We study these questions in a controlled setting where a compressed representation is decoded by a language readout, which we call the *linguistic readout* framework. The setting has three advantages: (i) the encoder bottleneck C directly controls superposition pressure ($\delta = 18/C$); (ii) the language task gives slot-level behavioral measurements (object slots c1/c2 vs. relation slot r) that localize failure and repair at sub-sentence granularity; (iii) the pipeline is fully reproducible (deterministic data generator, 5 seeds, pre-registered judgments with TOST $\delta = 0.05$).

Our experiments (E0–E22; 1000+ trained models, all pre-registered) establish the four results above, and we state our claims at two levels. In the controlled toy regime (one-hot inputs, linear-ReLU bottleneck), R1–R4 hold in full, including a continuous repair-locus transition ($\delta \approx 2.2$–2.6) that a Cover-capacity argument explains at first order. Across environments (toy, a semi-synthetic bridge, a small end-to-end CNN) a weaker but consistent set of regularities holds: failure is often an addressing mismatch rather than a total loss of information (the probing trap); repair is unit-local and dose-responsive with no cross-unit transfer; geometry that drifts globally does not equal repair; and repair lives in the readout or the encoder depending on compression pressure. The weak hypothesis does not promise a sharp transition band in arbitrary datasets or models. We contrast all of this with the geometric-repair hypothesis through pre-registered, decision-theoretic tests (R4).

2 Related Work

Compositional generalization and coverage. Empirical scaling analyses show that data coverage (not model family) drives visual compositional generalization (Uselis et al., 2025), and compositional risk minimization theory characterizes extrapolation from covered combinations (Mahajan et al., 2025). Our R2 replicates the coverage sufficiency result in a new setting. What we add is the *mechanism layer*: coverage works because readout pathways are rebuilt per unit (R2–R4), not because the representation is re-organized. This is complementary rather than competing: we answer *where* and *how* coverage repairs.

Superposition and binding. Superposition (Elhage et al., 2022) and the binding problem (von der Malsburg, 1981; Hinton, 1990) motivate why compressed representations might fail compositionally. We treat superposition as a *stress factor* (compression deepens failure, R1/R3) rather than a claim about superposition itself. Where the superposition line of work studies how and when representations collapse (Elhage et al., 2022), we study what happens *after* collapse: where repair lives and how compression regulates it. The new axis for this literature is that the *repair locus* depends on compression (R4).

Probing and causal interventions. Activation patching (Wang et al., 2023), causal tracing (Meng et al., 2022), and the logit lens (nostalgebraist, 2020) operate on weights/activations; linear probing has well-documented pitfalls (Alain & Bengio, 2017; Hewitt & Liang, 2019; Belinkov, 2022), and we make one consequence explicit (Lemma 1): a probe recovering information does not imply the decoder can *address* it, so probing alone cannot diagnose this class of failures. Our toolbox operates on the *task-level slot structure*: prefix injection $\times$ differential inversion (E3), slot-level probability and first-error diagnostics (E5–E7), and the frozen-encoder locus test (E9d). These tools complement weight-space methods: they answer *which slot-level pathway* breaks, while patching answers *through which internal path information flows*.

3 Experimental Setup

Toy regime. A scene is a pair of objects (A, B) with a relation $r \in \{\text{left_of}, \text{right_of}, \text{above}, \text{below}\}$; the scene is encoded as $f \in \mathbb{R}^{18}$ (one-hot: A color 4, A shape 3, B color 4, B shape 3, relation

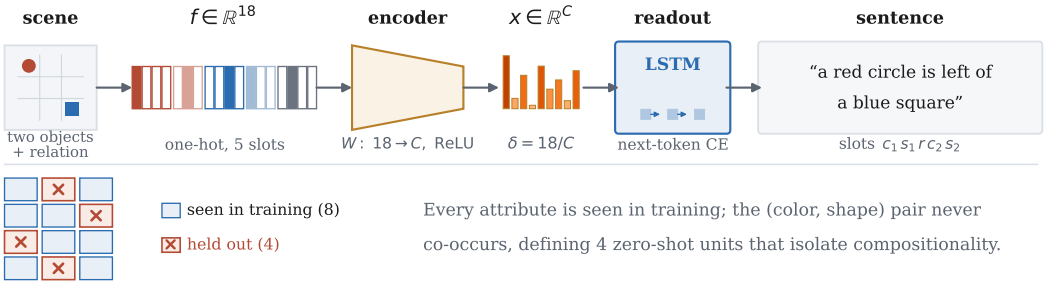


Figure 1: Setup. A scene is encoded to $f \in \mathbb{R}^{18}$, compressed by a trainable ReLU encoder to C channels ($\delta = 18/C$), and decoded by an LSTM into a five-slot sentence $c_1\, s_1\, r\, c_2\, s_2$. Four (color, shape) combinations are held out of training (lower band): every attribute still occurs there, but never in those pairs, giving a zero-shot compositional test.

4) and described by the sentence "a [c1] [s1] is [r] of a [c2] [s2]" (10 tokens incl. <eos>). A trainable linear-ReLU encoder compresses f to C channels ($\delta = 18/C$); an LSTM decoder reads $x = \mathrm{enc}(f)$ autoregressively. Compositional generalization: 12 (color, shape) combinations exist; 4 fixed combinations are unseen (never co-occur in training) while all single attributes are seen. Test-unseen contains 400 sentences per unseen combination (1,600 in the toy regime; the bridge uses 2,000 across 5 pairs). $C \in \{4, 6, 7, 8, 9, 10, 12, 18\}$ plus identity encoding (no compression) as reference (Fig. 1).

Semi-synthetic bridge. CIFAR-10 images (Krizhevsky, 2009) → frozen ResNet-18 GAP (512) (He et al., 2016) → PCA (64) → $f = [\text{c1 img}, \text{c2 img}, \text{relation}] \in \mathbb{R}^{132}$; same sentence template; 5 unseen (c1, c2) class pairs; LSTM C=64 ($\delta \approx 2.1$). Feature-class separability 0.854 bounds the task. This transfers the diagnostics to real visual features while keeping full control.

Protocol (S1). All claims: 5 seeds, medians with ranges, direction counts (≥4/5), and TOST ($\delta = 0.05$, bootstrap 90% CI) (Schuirmann, 1987; Lakens, 2017).[1] Judgments are reported per pre-registered rule without multiplicity correction, so we treat any single comparison as descriptive and rely on consistency across seeds and experiments for the main claims. Reported differences are differences of seed medians; medians of per-seed differences are given alongside for every key comparison (they can differ slightly under skew). All judgments were pre-registered: every E6–E9 design document fixes its decision rule before execution, and none were moved post hoc (disclosure table in §6).

Measures. Sentence accuracy (exact match); slot accuracy per position (SLOT_ATTR = {c1, s1, c2, s2} vs. SLOT_REL = {r}); per-unit aggregation (unit = unseen combination); and, for geometric analysis, encoder activation centers $\mu[c]$ = mean of $\mathrm{enc}(f)$ over test sentences of unit c, with displacement $\|\mu_m - \mu_0\|/\|\mu_0\|$ between a repaired model m and its same-seed baseline 0.

Formal setup. Let $f \in \{0,1\}^{18}$ be the scene feature, $x = \mathrm{ReLU}(Wf + b) \in \mathbb{R}^C$ the encoder output (compression $\delta = 18/C$), and

$$p_\theta(y_t \mid y_{<t}, x) = \mathrm{softmax}(Uh_t), \qquad h_t = \mathrm{LSTM}\big(h_{t-1}, [\,e(y_{t-1}); x\,]\big) \tag{1}$$

the serialized decoder with output layer U. A *unit* u (an unseen combination; not a network unit) is an unseen (color, shape) pair; $\mu(u) \in \mathbb{R}^C$ denotes its code centroid over scenes. We call the map $\pi_u : x \mapsto (\ell_1, \ldots, \ell_T)$ from the code to the slot logits of u's sentence the *readout pathway* of u; repair of u means rebuilding π_u so that $\arg\max \ell_t = y_t$ at every position.

4 Results

Figure 2 previews the full mechanism in schematic form: failure is a broken readout address (R1), augmentation rebuilds the combination's readout pathway locally (R2), and the repair locus shifts from the readout to the encoder as compression grows (R3/R4).

[1] One sub-setting (E8's toy identity encoding) used 3 seeds; no claim in this paper cites it. Every cited number is 5-seed.

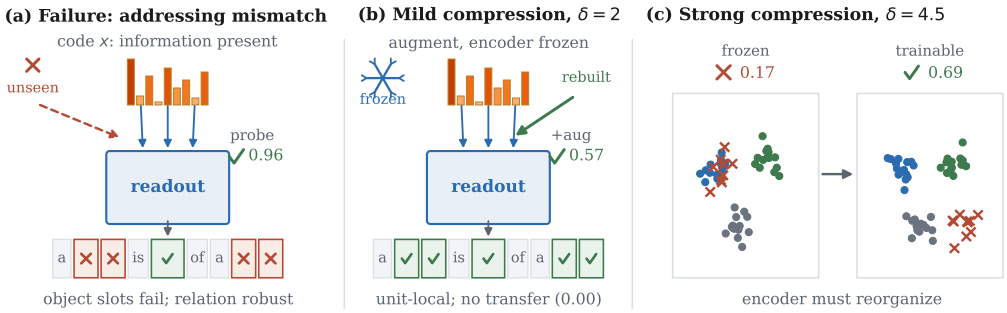


Figure 2: Mechanism. (a) Failure: the code retains the information (probe 0.96) but the readout pathway is broken, so object slots fail while the relation slot stays robust. (b) At δ=2 augmentation rebuilds the readout pathway unit-locally, with the encoder frozen (0.57) and no transfer to other units. (c) At δ=4.5 the unseen code is collapsed onto seen codes; repair requires encoder reorganization (frozen 0.17 versus trainable 0.69).

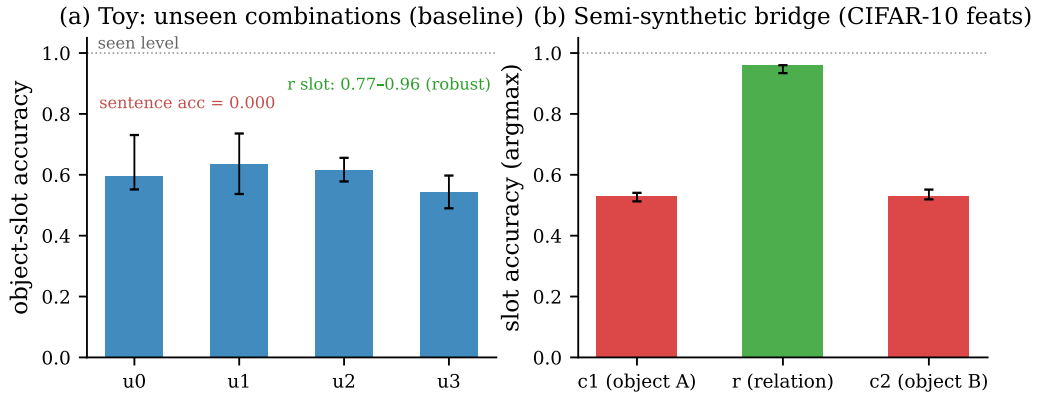


Figure 3: Failure diagnosis. (a) Toy: slot accuracy of the four unseen units at baseline (sentence accuracy 0.000; relation slot 0.77–0.96, above object slots 0.49–0.74). (b) Bridge: c1/r/c2 slot accuracies (LSTM C=64, 5 seeds, median $\pm$ range).

4.1 Failure is a content-addressing mismatch (R1)

Results in this section come from the toy regime first; the semi-synthetic bridge and the end-to-end CNN replicate the slot pattern (§4.7). Compositional generalization failure is not an information loss. The encoder retains the information (probe separability $\approx$ 0.96), but the decoder fails to *address* it for unseen combinations: object slots collapse toward chance while the relation slot remains substantially more robust.

Figure 3 shows this. In the toy regime, baseline unseen sentence accuracy is 0.000 (LSTM, C = 9) while per-unit slot accuracies (c1/s1/c2/s2, median over units) sit at 0.49–0.74 across the four unseen units and seeds (median 0.60); the relation slot is markedly higher, 0.77–0.96 (median 0.85), degraded relative to seen combinations but far above the object slots. The same pattern is sharper in the semi-synthetic bridge: c1 = 0.52, c2 = 0.55, r = 0.96 (5 seeds, LSTM C=64). Failure persists without compression (identity encoding, bridge: unseen 0.239 vs. seen 0.765; E7), so it is not an artifact of superposition.

Interpretation. The separation ratio $\approx$ 0.96 (E5b) shows information is recoverable; the slot pattern shows the *address* is missing. This locates failure at the readout layer, which is the first clue that repair may also live there.

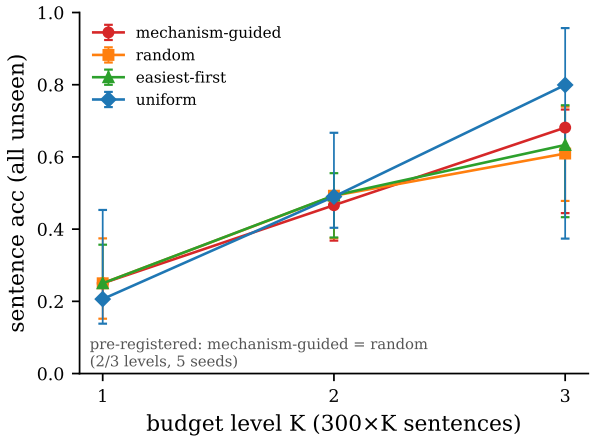


Figure 4: Budget–effect curves (toy, C=9, 5 seeds, median ± range). Mechanism-guided selection is not distinguishable from random at any level (pre-registered P1: 0/3); the four strategies separate only weakly in the toy regime, while the bridge shows uniform ≫ mechanism-guided (§4.2).

Excluding alternatives (E14, E19). Two accounts other than addressing must be ruled out. Capacity is not the limit: enlarging the decoder hidden size from 128 to 512 and 1024 (77K to 4.2M parameters) leaves unseen sentence accuracy essentially unchanged (0.000 at both C=4 and C=9 for hidden 128 and 512; 0.006 and 0.036 at 1024), and on the bridge the same holds (hidden 64→256: 0.133→0.105, slot pattern unchanged, E19). Exposure bias is not the driver either: under teacher forcing with gold prefixes, so that no error accumulates, the object slots reach only 0.57–0.61 while the relation slot reaches 0.976–0.998. The decoder recovers the relation almost perfectly yet cannot select the object attributes with the correct context in hand: the signature of a missing address, not a missing capacity.

4.2 Repair is coverage-driven, unit-local, and thresholded (R2)

Given a fixed augmentation budget, selection method does not matter in the toy regime: mechanism-guided selection (prioritizing the most address-mismatched units) is never better than random. In the bridge, uniform coverage across all unseen units dominates mechanism-guided selection. Repair of a covered unit follows a steep dose–response in per-unit exposure rather than a step function, and repair does not transfer to uncovered units.

Budget–effect curves (Fig. 4; toy, C=9, 5 seeds, medians; total budget $300 \times K$ sentences) support this. Sentence accuracy across *all* unseen units under uniform coverage rises $0.206 \rightarrow 0.490 \rightarrow 0.799$ for budget levels $K = 1 \rightarrow 2 \rightarrow 3$. Mechanism-guided augmentation (prioritizing the units with the highest pre-registered addressing-mismatch score) is *not* better than random at any level (pre-registered P1: 0/3 levels; K=2 mech 0.466 vs. rand 0.492, 1/5 seeds; K=3 +0.072, 3/5). In the bridge, uniform (0.703/0.763) ≫ random (0.389/0.501) > mechanism-guided (0.351/0.477); the P3 pre-registered verdict is that coverage wins and mechanism-guided is significantly *worse* (CI entirely negative, 2/2 levels).

Unit-level structure sharpens the picture. Under focused augmentation (mechanism-guided, K=2, 300 sentences/unit), *covered* units repair to 0.70–1.00 (median 0.93) while *uncovered* units stay at baseline (0.00–0.21; sentence accuracy ≈ 0.25, range 0.15–0.37, across the focused strategies, consistent with roughly one unit repaired in four). No transfer: pre-registered P2 shows no slot improvement in uncovered units (median -0.012/-0.005, 1/5 seeds), so a pathway is rebuilt only when that combination is augmented.

Per-unit exposure, manipulated orthogonally (E15). The comparison above confounds per-unit count with the number of covered units, so we fix coverage at all four units and vary only the per-unit sentence count (pre-registered in E15_DESIGN). Median accuracy rises monotonically (0 / 75 / 150 / 300 / 600 sentences per unit: $0.004 \rightarrow 0.224 \rightarrow 0.641 \rightarrow 0.871 \rightarrow 0.974$), with per-seed Spearman $\rho = 1.0$ at 5/5 seeds and a 300-vs-75 gap of $+0.646$. Gains per doubling decline ($+0.42$,

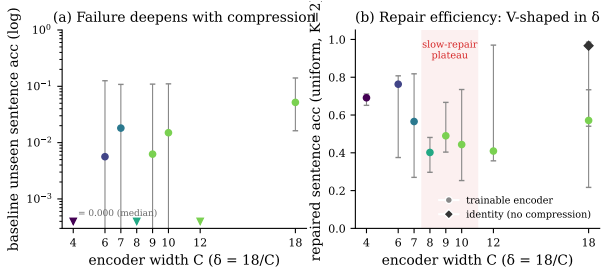


Figure 5: Compression regulation. (a) Failure depth increases with compression at the scan endpoints (log scale; 8 points, median ± range; interior 0.000–0.018 fluctuations). (b) Repair efficiency is non-monotone in δ; identity (diamond) repairs fastest. Shaded band: slow-repair plateau $\delta \in [1.5, 2.25]$.

$+0.23$, $+0.10$), so repair saturates rather than switching at a single point: "thresholded" here means a steep dose–response, not a step function. This replicates the coverage result (Uselis et al., 2025; Mahajan et al., 2025); what it does not predict is the *unit-locality and this dose–response*.

4.3 Compression deepens failure and shapes repair efficiency (R3)

This section is specific to the toy regime. Compression deepens failure and shapes repair efficiency non-monotonically, with mid-compression ($\delta \approx 1.5$–2.25) repairing slowest.

Figure 5 shows the scan. Baseline unseen sentence accuracy (medians) at the endpoints: $C{=}4 \to 0.000$, $C{=}9 \to 0.006$, $C{=}18 \to 0.052$, identity $\to 0.195$; failure deepens with compression at the endpoints. The interior scan ($C{=}6$–12) shows small non-monotone fluctuations (0.000–0.018, with $C{=}8$ and $C{=}12$ at 0.000), so we claim the endpoint trend, not strict monotonicity. Under uniform repair ($K{=}2$, 150 sentences/unit): $C{=}4 \to 0.691$, $C{=}6 \to 0.763$, $C{=}9 \to 0.490$, $C{=}18 \to 0.571$, identity $\to$ 0.967. The slow-repair plateau spans $\delta \in [1.5, 2.25]$ ($C{=}8$: 0.402, $C{=}12$: 0.409). A caution: the $C{=}6$ point (0.763) deviates from a clean valley, so we claim the shape only as an observation across the scan.

Confound control. A trainable $C{=}18$ encoder (0.571) underperforms the fixed identity encoder (0.967) by 0.396 (CI $[-0.523, -0.104]$); trainability itself is a confound. All within-trainable comparisons (the V-shape) are unaffected; comparisons to identity are reported with this caveat.

4.4 Representation geometry changes globally; behavior repairs locally (R4, observation)

Augmentation re-organizes encoder geometry globally, including for units that are never augmented, yet behavior repairs only locally. The behavioral slot measurement reveals a locality that geometric analysis cannot.

Figure 6 shows this. For focused (mechanism-guided, $K \in \{1, 2\}$) repair at every compression level $C \in \{4, 9, 18\}$, the ratio of uncovered/covered median activation displacement (rel. L2) is ≥ 0.5 in 30/30 model-seeds at $K \in \{1, 2\}$: uncovered units' encoder activations move as much as, sometimes more than, covered ones (median ratio 1.04 at $C{=}4$, 0.98 at $C{=}9$, 1.07 at $C{=}18$). Extending to $K{=}3$ yields 44/45. Yet the behavior of uncovered units is unchanged (sentence accuracy 0.00–0.21; P2 no transfer). Displacement magnitudes follow the repair-efficiency order (Spearman $\rho = 0.48$). On the bridge the dissociation replicates (E18): with augmentation focused on one of five unseen pairs, the covered pair repairs (0.09 $\to$ 0.88) while the other four stay within 0.02 of baseline, yet their activation centers move 0.81–1.03$\times$ as far as the covered pair's (median 0.91, 5/5 seeds).

Interpretation. Geometric change is incidental (the encoder drifts), not the mechanism of repair, at least at mild compression; the mechanism is local to the readout pathway, which geometric analysis alone cannot see. Slot-level behavior resolves locality that activation geometry obscures. The linguistic readout is what makes this visible.

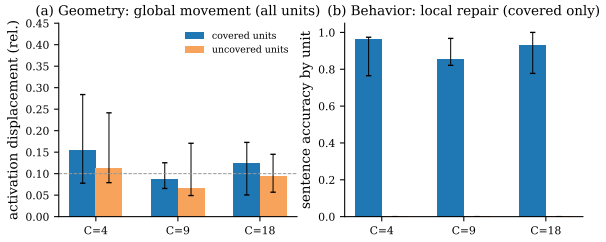


Figure 6: Geometry–behavior dissociation. (a) Activation displacement (rel. L2) is global: uncovered units move as much as covered units at every C. (b) Behavior is local: only covered units repair. Same models, both measurements.

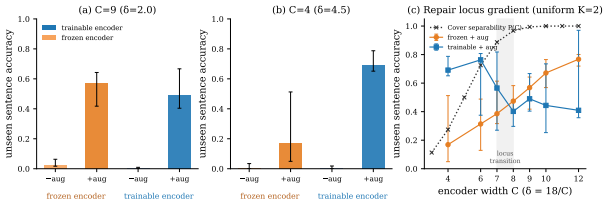


Figure 7: Frozen-encoder locus test (5 seeds, median $\pm$ range). (a) At C=9 the readout pathway alone is sufficient (frozen $\approx$ trainable with augmentation); (b) at C=4 encoder reorganization is required (frozen $\ll$ trainable). (c) Seven-point scan: frozen repair improves monotonically with C while trainable repair traces the non-monotone plateau; the locus transition band lies between C=7 and C=8 ($\delta \in [2.2, 2.6]$). The Cover-theory prediction applies to the toy one-hot setting, not to continuous visual features.

4.5 The repair locus shifts with compression (R4, causal test)

This section is specific to the toy regime; §4.7 tests the weak form in more realistic settings. At $\delta = 2$ the readout pathway is the *sufficient* repair site: freezing the encoder (training only the decoder) still repairs covered units, with no loss. At $\delta = 4.5$, the encoder must participate, since freezing costs 0.52. The repair locus thus *shifts* with compression.

A pre-registered 2×2 factor experiment (Fig. 7; frozen/trainable $\times$ augmented/not, $C = 9$ and $C = 4$, 5 seeds, same-run comparisons) gives:

- $C = 9$ ($\delta = 2.0$): frozen+aug repairs covered units (median 0.568 vs. 0.026 without aug; difference of seed medians $\Delta = +0.54$, median of per-seed differences $+0.51$, 5/5 seeds positive, bootstrap CI $[+0.444, +0.567]$); frozen is *not* worse than trainable (0.568 vs. 0.490; $+0.078$ as difference of medians, $+0.031$ as median of per-seed differences, 3/5 seeds positive, max per-seed deficit 0.073). Slot pattern of repair preserved.
- $C = 4$ ($\delta = 4.5$): frozen+aug repairs weakly (0.169 vs. 0.000; $\Delta = +0.17$, 5/5) and is *much* worse than trainable (0.169 vs. 0.691; -0.52 as difference of medians, -0.57 as median of per-seed differences, 0/5, CI $[-0.619, -0.370]$).

Locus gradient (E9e, seven-point scan, Fig. 7c). Extending the frozen experiment to $C \in \{6, 7, 8, 10, 12\}$ (100 additional models) shows the shift is continuous: frz−tr rises monotonically (-0.522 at C=4 $\rightarrow$ -0.180 at C=7 $\rightarrow$ $+0.072$ at C=8 $\rightarrow$ $+0.358$ at C=12), with the transition band between C=7 ($\delta = 2.6$) and C=8 ($\delta = 2.2$). Two structures emerge. Frozen repair improves monotonically with C (0.169 $\rightarrow$ 0.767) while trainable repair traces the non-monotone plateau of §4.3: the plateau disappears when the encoder is frozen, so it is a property of encoder reorganization, not of the readout. R3 and R4 thereby unify: at mild compression the readout repairs alone and encoder drift is a byproduct (at C=10/12 actively harmful, frozen better by $+0.228/+0.358$; this overtaking regime is LSTM-specific, §4.7); in the plateau band reorganization is slow and unreliable; at strong compression it is necessary and efficient.

Cross-validation with §4.4: displacement is necessary at C=4 (frozen fails) but incidental at C=9 (frozen succeeds).

4.6 A FIRST-ORDER ACCOUNT OF LOCALITY AND LOCUS

The empirical chain above admits a compact first-order explanation. We state three results; proofs and caveats are deferred to Appendix D.

Lemma 1 (probes cannot see the failure). *In a ReLU linear region, the function class of a linear probe on $x = Wf$ is exactly the row space of W: a slot label $z = \beta^\top f$ remains linearly decodable from x iff $\beta \in \mathrm{row}(W)$, with min-norm probe $\theta = (W^\top)^\dagger \beta$.* Hence linear-probe accuracy is invariant under the encoder's linear mixing: separability ≈ 0.96 is consistent with *any* level of behavioral failure (R1). Probes measure information presence, not addressability.

Proposition 2 (first-order locality and threshold). *One SGD step on the decoder's output layer with a batch from unit u gives $\Delta U = \eta \sum_t (e_{y_t} - p_t)\, h_t^\top$; the induced logit change on another unit u' is*

$$\Delta \ell(u') \;=\; \eta \sum_t (e_{y_t} - p_t)\, \big\langle h_t(u),\, h_t(u') \big\rangle. \tag{2}$$

When per-unit hidden trajectories are near-orthogonal, first-order transfer vanishes, so repair is unit-local (R2, no transfer). The same expansion yields a threshold: flipping the argmax requires the accumulated margin to exceed the logit gap γ, i.e. $n \gtrsim \gamma/(\eta s)$ examples at per-example signal s. Repair therefore appears all-or-nothing rather than graded.

Proposition 3 (the locus transition is a Cover-capacity transition). Addressing unit u at the output layer requires an affine hyperplane separating its code from the 11 remaining unit codes. By Cover's function-counting theorem (Cover, 1965), for 12 codes in general position in $\mathbb{R}^C$, the fraction of implementable dichotomies (hence the probability that a generic unseen code admits a readout pathway) is

$$P(C) \;=\; 2^{-11} \sum_{k=0}^{C} \binom{11}{k} \;=\; 0.27,\ 0.50,\ 0.73,\ 0.89,\ 0.97,\ 0.99 \quad (C = 4, 5, 6, 7, 8, 9). \tag{3}$$

At $C{=}4$ the frozen code generically lies inside the convex hull of the seen codes: no hyperplane addresses it, and the only way to create a pathway is to move the code to a hull vertex. Encoder reorganization *is* the repair. At $C \geq 8$ the pathway exists with probability ≈ 1, and decoder-only repair suffices. The observed transition band $C \in (7, 8)$ coincides with $P(C)$ crossing ≈ 0.9 (Fig. 7c, dotted curve).

An independent geometric test. Aligning a transition point with a monotone curve is a weak test, so we measure the geometry directly: ρ, the fraction of an unseen unit's codes lying inside the convex hull of the seen codes (no hyperplane separates such a code). Measured ρ falls from 0.40 at $C{=}4$ to 0.00 by $C{=}8$, tracking $1 - P(C)$ in direction and below the bound, as ReLU folding predicts. Corroboration from a second observable, not a fitted threshold; ρ is a compression-level switch, not a unit-level predictor. Post-hoc.

Caveats. The account is first-order: codes are learned, not in general position (ReLU folding lowers separability, so $P(C)$ is an upper bound), and the LSTM poses a per-step hyperplane problem. Its value is that it *predicts the transition location from first principles*, with no fitted parameter.

4.7 WEAK-FORM VALIDATION IN SEMI-SYNTHETIC AND END-TO-END CNN SETTINGS

Can the locus mechanism be converted into a training rule? The natural candidate is *freeze the encoder and fine-tune only the decoder with augmentation*. We test it on the semi-synthetic bridge ($C{=}64$, $\delta \approx 2.06$, on the mild side of the toy transition band).

Result 1 (negative): the toy locus does not transfer naively. A randomly initialized, frozen encoder fails: decoder-only training reaches 0.192 versus 0.753 for full fine-tuning (-0.561; 0/5 seeds; E10). Continuous image features do not admit readout-only repair from a random code (toy one-hot codes stay near-orthogonal after mixing, Prop. 3); the locus transition is thus a property of the toy code geometry.

Result 2 (positive, weak form): freezing a *trained* encoder recovers most of the repair. When the encoder is trained on the base task and then frozen, decoder-only augmentation fine-tuning reaches 0.623 (0.605–0.687, 5/5 seeds) versus 0.753 for full fine-tuning (E11): the decoder update alone

accounts for ≈79% of the repair gain. A bridge compression scan (E16, Appendix A-2) shows the frozen deficit is a flat offset, -0.12 to -0.16 at every C, with the recovery rate stable at 75–81%.

Extension to end-to-end training (E12). The bridge above uses a frozen visual front-end (ResNet-18 + PCA). We repeat the test end to end: a trainable SmallCNN (0.3M parameters) replaces the frozen features and is trained jointly with the LSTM (C=64, $\delta \approx 2.06$, 5 seeds, pre-registered in E12_DESIGN). Two things carry over. The failure pattern reproduces: baseline unseen accuracy is 0.020 (0.003–0.025) with object slots at 0.392 and the relation slot at 0.999, so R1 holds under a fully learned encoder. And the weak form gets *stronger*: full fine-tuning repairs to 0.606, while freezing the trained CNN and fine-tuning the decoder alone reaches 0.539 (0.519–0.578), recovering 88.5% of the repair gain (frz−full -0.066, 0/5). A learned code is more amenable to readout-only repair than a PCA code.

Decoder architecture (E20). Repeating the toy locus test with a small Transformer decoder (2 layers, 4 heads, d=64) transfers the failure pattern (object slots 0.65 vs. relation 0.96 at C=9) and the weak form, but not the overtaking regime: frozen repair recovers 82% of the gain at C=9 and 90% at C=12, while full training is never worse at any C, so freezing *helps* only under the LSTM. The recovery rate rises with C (51% at C=4): the weak-form gradient that the bridge did not show (Appendix A-5).

Takeaway. The practical rule is a quantified trade-off, not a free lunch: freezing the encoder and fine-tuning the decoder with augmentation recovers 79% of the repair gain on the bridge, 88.5% end to end, and 82–90% with a Transformer decoder. The locus shift itself stays bounded to the toy regime.

5 Discussion

What augmentation repairs. Coverage works because readout pathways are rebuilt per unit: unit-locally (no transfer), thresholded (≈150–300 sentences/unit), and coverage-driven (selection method irrelevant beyond coverage). This is consistent with, and mechanistically completes, the coverage sufficiency result.

Where augmentation repairs. The readout pathway at mild compression; the encoder at strong compression. Compression is not merely a *deepener* of failure (R1/R3) but a *switch* for the repair locus (R4); the geometry-behavior dissociation warns that activation drift can be incidental.

The measurement contribution. Slot-level behavioral measurement resolves what activation geometry cannot (Fig. 6): the locality of repair. The toolbox (prefix injection × differential inversion, slot diagnostics, frozen-encoder locus test) is toy-independent and released as a reusable package.

Limitations. (i) The sharp locus-shift transition holds only in the toy one-hot regime with a linear-ReLU bottleneck; in continuous image features only the weak form reproduces, with no clear transition band. (ii) All tasks use fixed-template generation, with no free text or nested recursion. (iii) Decoder architecture is tested in the toy: a small Transformer transfers the failure pattern and the weak form (§4.7); on COGS the slot-selectivity pattern transfers but only in a partially-learned regime (gen exact-match 0.000), and the probing trap and freezing gradient await a compositional-augmentation protocol (Appendix A-6). (iv) We claim no quantitative threshold transfer; δ is toy-internal and real models have no single bottleneck C. (v) Future work: test the weak hypothesis on COGS and SCAN with a compositional-augmentation protocol, and define an equivalent compression-pressure measure for real models.

6 Conclusion

Augmentation repairs compositional generalization by rebuilding readout pathways, unit-locally and thresholded. Compression moves the repair locus: at $\delta = 2$ the readout alone suffices, at $\delta = 4.5$ the encoder must participate.

AI USE STATEMENT

In this work, we used generative AI tools for experiment design discussion, code implementation, log summarization, and manuscript drafting. We have not used generative AI tools for generating hypotheses, for executing pre-registered decision rules, or for producing any reported number, and the remaining disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for copy-editing and for the layout of several figures. We have reviewed all AI-assisted work: every reported number was recomputed from raw experiment logs by an independent verification script, all pre-registered decision rules were fixed before execution, and every figure was regenerated from stored result files rather than from AI output. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses only synthetic data and CIFAR-10, involves no human subjects and no personal or sensitive data, and raises no foreseeable harmful applications.

REPRODUCIBILITY STATEMENT

All experiments are deterministic given the seed. The anonymized supplementary material contains the data generator, training code, measurement and figure scripts, pre-registration design documents (E0–E22), and raw result files. Each toy training run takes about 14 seconds on a single consumer GPU (RTX 4060, 8 GB) and each end-to-end CNN run about 20 seconds; the full experiment matrix (about 1000 models) requires roughly 6 GPU-hours. A number-auditing script re-derives every value reported here from the raw logs.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *ICLR Workshop*, 2017.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Thomas M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, EC-14(3):326–334, 1965.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. Transformer Circuits Thread, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *EMNLP*, pp. 2733–2743, 2019.

Geoffrey E. Hinton. Mapping part–whole hierarchies into connectionist networks. *Artificial Intelligence*, 46(1-2):47–75, 1990.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang, Marc van Zee, and Olivier Bousquet. Measuring compositional generalization: A comprehensive method on realistic data. In *ICLR*, 2020.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Brenden M. Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *ICML*, pp. 2873–2882, 2018.

Daniel Lakens. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362, 2017.

Divyat Mahajan, Mohammad Pezeshki, Charles Arnal, Ioannis Mitliagkas, Kartik Ahuja, and Pascal Vincent. Compositional risk minimization. In *ICML*, 2025.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *NeurIPS*, volume 35, 2022.

nostalgebraist. Interpreting gpt: The logit lens. LessWrong, 2020.

Laura Ruis, Jacob Andreas, Marco Baroni, Diane Bouchacourt, and Brenden M. Lake. A benchmark for systematic generalization in grounded language understanding. In *NeurIPS*, volume 33, 2020.

Donald J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680, 1987.

Arnas Uselis, Andrea Dittadi, and Seong Joon Oh. Does data scaling lead to visual compositional generalization? In *ICML*, 2025.

Christoph von der Malsburg. The correlation theory of brain function. Technical report, MPI Biophysical Chemistry Internal Report 81-2, 1981.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in gpt-2 small. In *ICLR*, 2023.

A PRE-REGISTRATION DISCLOSURE

Decision point	When fixed	Outcome / disclosure
S1 protocol upgrade 2→5 seeds + TOST	before E6 reruns	response to review; disclosed
E7 R3 gate (absolute 0.95 → relative unseen≈seen)	after results	design flaw (feature ceiling 0.854), not selection; disclosed
E8 P3 TOST semantics (CI one-sided = "worse $< \delta$")	at judgment	wording corrected to CI substance; direction unchanged
E9a A1 direction semantics (script bug)	at summary	fixed before any claim; corrected verdict (reversed)
E9d E9d2 operationalization ("not worse" $\neq$ "better")	at summary	clarified; direction unchanged
E9e intermediate-compression frozen scan (E9d extension)	executed after E9d	follow-up motivated by Limitation (ii); decision rules and protocol inherited verbatim from E9d, disclosed as not independently pre-registered
E10/E11 bridge freezing tests (practical-rule question)	executed after E9e	E10 pre-registered in E10_DESIGN; E11 remedial weak-form follow-up after E10's negative result, both disclosed in §4.7
E16 bridge compression scan ($C \in \{32, 48, 96, 128\}$, weak-form gradient test)	executed after E12	trend predicted in the A-2 plan before execution; prediction (frozen deficit grows with compression) not confirmed, flat offset reported
E18 bridge focused-augmentation dissociation test (A-3)	executed after E16	pre-registered in E18_DESIGN (covered pair fixed, both judgments fixed before execution); P1 and P2 passed
E19 bridge capacity control (A-4)	executed after E16	pre-registered in E19_DESIGN; judgment reported in Appendix
E17 toy transition ablations (noise, MLP encoder; A-1)	executed after E9e	sensitivity analysis inheriting the E9e protocol; not an independent pre-registered design, reported as post-hoc in Appendix C
E20 Transformer-decoder replication (B-2)	executed after E18/E19	pre-registered in E20_DESIGN; P1/P3 passed, P2 (locus overtaking) not reproduced, reported honestly
E21/E22 COGS weak-hypothesis test (B-1)	executed after E20	pre-registered in E21_DESIGN; no-attention version did not transfer, attention version transfers slot-selectivity (rel>obj) but P1/P3 need a compositional-augmentation protocol, reported in Appendix A-6
All E6–E9 decision rules	before each execution	fixed; none moved post hoc

Table 1: Pre-registration disclosure: every decision-point modification, its timing, and its nature.

B BROADER IMPLICATIONS

Our results come from a controlled regime, so we state the transferable part carefully: what carries over is the *diagnostic and decision framework*, not the specific numbers. Three implications follow with different levels of support.

Addressing versus missing knowledge (best supported). R1 and E14 show that failure can coexist with recoverable information (probe separability 0.96 against 0.00 sentence accuracy), that it concentrates in entity-like slots while relational slots stay near ceiling (0.998), and that neither decoder capacity (a 50× parameter increase) nor exposure bias (teacher forcing with gold prefixes) accounts for it. Failure modes with this signature in generative models are natural candidates for the same diagnosis: factual hallucination in language models, and attribute-binding failures in text-to-image models, where the model demonstrably "knows" the concepts but binds them to the wrong slot. If the diagnosis holds, the remedy is targeted coverage of the failing units rather than more data in general.

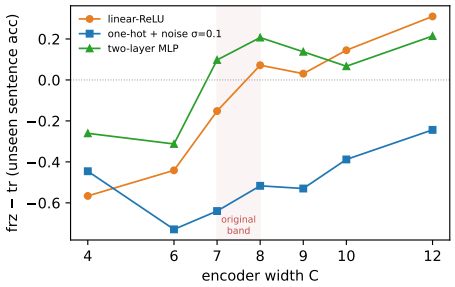


Figure 8: Locus-transition sensitivity (E17, 5 seeds, medians of per-seed differences). Input noise (σ=0.1) removes the crossover entirely (frz−tr negative at every C); a two-layer MLP encoder keeps a crossover but earlier (C=6–7) and weaker, with an unstable trainable baseline.

Freezing as a conditional rule (supported end to end). E10 versus E11 and E12 give a conditional rule rather than a universal one: freezing a randomly initialised encoder fails outright (0.192 against 0.753), whereas freezing an encoder after it has been trained recovers most of the repair gain (79% on the bridge, 88.5% end to end) at no encoder update. Multimodal pipelines that pretrain an alignment stage and then freeze the vision tower for instruction tuning match the second condition; pipelines that freeze an encoder before it has been trained match the first and should not expect the saving. Compression severity sets the boundary: at strong compression the locus moves to the encoder and freezing caps performance.

What we would not extrapolate. The per-unit counts (75/150/300/600) and the slow-repair region are properties of this task, vocabulary, and model size; we expect the shape (a steep, saturating dose–response) to transfer more reliably than the location of the knee. The Cover-capacity formula assumes codes in general position, which trained high-dimensional representations violate, so $P(C)$ is an upper bound and the agreement in Fig. 7c is corroboration rather than a fitted threshold. Testing each of these predictions at scale is the natural next step: whether per-unit exposure thresholds exist for factual recall, and whether the freeze/train boundary tracks projection dimensionality in vision-language models. Each is a direct falsification test of the mechanism proposed here.

C Ablations and robustness (A-layer)

This section collects the four robustness experiments that bound how far the toy-regime results carry. Each is reported at face value, including the negative ones.

A-1: the transition is a one-hot artifact, and this is consistent with our boundary claims (E17). We ask how the sharp locus transition depends on design choices, by repeating the frozen-encoder scan under two perturbations: (a) adding Gaussian noise ($\sigma = 0.1$) to the one-hot inputs, and (b) replacing the linear-ReLU encoder with a two-layer MLP. With input noise the transition disappears: the frozen deficit frz−tr stays negative across the whole scan (-0.73 at C=6 to -0.24 at C=12; 0/5 at every C), so frozen repair never overtakes full training. The MLP encoder keeps the qualitative phenomenon but earlier, weaker, and noisier: frz−tr crosses zero between C=6 and C=7 (span $+0.48$ versus $+0.88$ for the linear encoder), with an unstable trainable baseline (Fig. 8). Together these pin down why real continuous features show no sharp band: the transition rests on near-orthogonal one-hot codes (Proposition 3), and mild input noise erases the separability structure it needs. This is an internal explanation for the external-validity boundary we report in §4.7, not a new extrapolation claim.

A-2: bridge compression scan (E16). On the semi-synthetic bridge we swept the PCA bottleneck $C \in \{32, 48, 64, 96, 128\}$ (5 seeds, frozen/trainable × augmentation; C=64 from E11). The frozen deficit is a flat offset rather than a function of compression: frz−tr lies between -0.12 and -0.16 at every C (0/5 seeds at each), with no monotone trend across a 4× range. The recovery rate of freezing a trained encoder is correspondingly stable at 75–81%. Baseline unseen accuracy deepens mildly with compression (0.095 at C=32 to 0.158 at C=128), consistent in direction with R3. We note

that the bridge decoder also receives $132-$C uncompressed spatial/relation features, so the effective compression is smaller than the nominal δ.

A-3: geometry–behavior dissociation on the bridge (E18). We repeat the §4.4 dissociation with continuous features: augmentation focused on one of the five unseen pairs (the pair fixed in advance). Behavior is strictly local: the covered pair repairs from 0.09 to 0.88 while the other four stay within 0.02 of baseline (5/5 seeds). Geometry is global: the uncovered pairs' activation centers move 0.81–1.03$\times$ as far as the covered pair's (median 0.91, 5/5 seeds). Both pre-registered judgments pass.

A-4: bridge capacity control (E19). Enlarging the LSTM hidden size from 64 to 256 (about 5.6$\times$ parameters) leaves unseen accuracy essentially unchanged (0.133 to 0.105) and the slot pattern intact (rel$-$attr $\approx +0.43$ at every size). Combined with E14, the capacity and exposure-bias exclusions now hold in two environments.

A-5: decoder architecture (E20). Replacing the LSTM with a small Transformer decoder (2 layers, 4 heads, d=64) and repeating the toy locus scan gives three findings. (i) The failure pattern transfers: at C=9 baseline unseen accuracy is 0.041 with object slots at 0.67 and the relation slot at 0.96. (ii) The weak form transfers with a visible gradient: the recovery rate of frozen-encoder repair rises from 51% at C=4 to 82% at C=9 and 90% at C=12 (median across 5 seeds; 0.68/0.71 at C=7/8). (iii) The overtaking regime does not transfer: frz$-$tr stays negative at every C (-0.08 to -0.40), so full training is never worse than freezing, and the mild-compression regime where freezing *helps* (§4.5) is LSTM-specific. The Transformer decoder is also uniformly stronger (full training reaches 0.81–0.99 at every C).

A-6: a COGS probe (E21, E22). We tested the weak hypotheses on COGS. A no-attention version of the code framework did not transfer: the decoder never learned the base task (exact-match 0.000), because a single code vector loses entity information. Adding attention to the decoder yields only partial, token-level learnability on the generalization split: token accuracy rises to 0.66 at the pre-trained checkpoint (0.61–0.62 after adaptation) versus 0.166 without attention, while exact-match accuracy on that split stays at 0.000 in all 18 runs. COGS is therefore far from solved in our framework, and we claim no improvement over the COGS literature; the value of the probe is diagnostic, not competitive. Within this partial regime the slot-selectivity pattern does transfer: structural tokens remain more learnable than entity tokens by a consistent margin (rel $-$ obj $= +0.14$ at C=64 and $+0.12$ at C=256, 3/3 seeds at each C), echoing R1's object-slot collapse. The probing trap and the freezing gradient, however, are not yet testable there: with attention the code vector is no longer the sole information channel (a linear probe on it recovers only 0.03–0.10), and the freezing test needs a compositional-augmentation protocol rather than an in-distribution fine-tuning slice, which we leave to future work.

D Proofs and caveats for §4.6

Proof of Lemma 1. Inside a ReLU linear region, $x = Wf + b$ holds exactly (all pre-activations positive), so a probe computes $\theta^\top x = (W^\top\theta)^\top f + \theta^\top b$. The probe can therefore implement exactly the linear functionals $\{\beta^\top f : \beta \in \mathrm{col}(W^\top)\} = \mathrm{row}(W)$. The system $W^\top\theta = \beta$ is solvable iff $\beta \in \mathrm{row}(W)$, and the minimum-norm solution is $\theta = (W^\top)^\dagger\beta$. Across regions the statement holds piecewise; empirically the slot labels of our task are decodable to ≈ 0.96, i.e. they lie approximately in $\mathrm{row}(W)$. □

Remark. Lemma 1 formalizes why probe separability cannot diagnose the R1 failure: it certifies information *presence* in the code's row space, not the existence of a trained readout pathway π_u. This is the precise sense in which our slot-level behavioral measure is strictly more informative than probing for this question.

Proof of Proposition 2. For cross-entropy loss on a batch from unit u, $\partial\mathcal{L}/\partial U = -\sum_t (e_{y_t} - p_t)h_t^\top$, giving $\Delta U = \eta\sum_t (e_{y_t} - p_t)h_t^\top$ for one SGD step. Evaluating the perturbed logits on unit u': $\Delta\ell(u') = \Delta U\, h(u') = \eta\sum_t (e_{y_t} - p_t)\langle h_t(u), h(u')\rangle$. If $\langle h_t(u), h(u')\rangle \approx 0$ for $u' \neq u$ (near-orthogonal per-unit hidden trajectories), the first-order transfer vanishes; the residual transfer is $O(\eta^2)$ plus trajectory drift. For the threshold, let γ be the logit gap between the correct token and the runner-up at the decision position; after n examples with aligned per-example margin signal s, the accumulated margin is $\approx \eta n s$, and the argmax flips when $\eta n s \gtrsim \gamma$, i.e. $n \gtrsim \gamma/(\eta s)$. Below

threshold the accuracy gain is exponentially small in the remaining gap (softmax), producing the observed all-or-nothing behavior. □

Proof of Proposition 3. A unit code $\mu(u)$ is addressable by an affine hyperplane iff it does not lie in the convex hull of the remaining 11 codes (strict separation of a point from a finite set). Cover's function-counting theorem (Cover, 1965) counts the dichotomies of N points in general position in $\mathbb{R}^C$ realizable by affine hyperplanes as $2\sum_{k=0}^{C}\binom{N-1}{k}$; dividing by 2^N dichotomies with $N{=}12$ gives $P(C) = 2^{-11}\sum_{k=0}^{C}\binom{11}{k}$. When $\mu(u)$ lies inside the hull, no output-layer pathway exists for u; creating one requires moving $\mu(u)$ to a vertex of the joint hull, an encoder-side change (reorganization). When $\mu(u)$ is a vertex, a pathway exists and decoder-only training can instantiate it. □

Caveats. (i) Codes are learned, not sampled in general position; ReLU folding can only decrease effective separability, so $P(C)$ is an upper-bound guide. (ii) Unit codes vary with the relation slot; the 12 points are centroids over the object-identity factor. (iii) The LSTM conditions each step on h_{t-1}, so the effective per-step problem is a sequence of hyperplane problems in the hidden space, not a single one; the agreement of the predicted and observed transition bands (Fig. 7c) is evidence that the first-order picture captures the dominant constraint.

E FULL JUDGMENT TABLES

Complete per-seed values for every pre-registered judgment (E6 A–D, E7 R1–R4, E8 P1–P3, E9a A1–B, E9b M1, E9c E9c1–2, E9d E9d1–3) are provided in the supplementary repository as machine-readable JSON together with the fixed decision rules.