SAME FIT, DIFFERENT FUNCTIONS: EVIDENCE STRUCTURE AND TRAINING ORDER SHAPE HELD-OUT BEHAVIOR

Anonymous authors
Paper under double-blind review

ABSTRACT

Training fit alone does not determine held-out behavior. We study controlled contradictory-supervision tasks whose runs reach near-ceiling or exact terminal training classification accuracy but express different held-out mappings. In Study I, Qwen2.5-0.5B is trained on a fixed budget of 200 exceptions. Increasing the fraction aligned with a shared equality relation raises mean held-out transfer from 0.0808 to 1.0000 across five coherence levels, strictly in all three seeds. A separate complete-cell holdout preserves the endpoint contrast, with a mean gain of 0.9000; its descriptive five-level extension is strictly increasing in every seed. A modular-composition family reproduces the positive fully coherent versus fully idiosyncratic endpoint contrast in three fresh seeds, but not the graded response. Study II first establishes a primary predictor. All 21 Phase-2 arms fit their training examples, yet 200 fully coherent corrections yield primary retention $P = 1.000$ and secondary transfer $G = 1.000$ in every seed, whereas 100 yield $G = 0.0125, 0.7825, 0.0300$. In an exploratory, post-hoc fixed-source intervention, six conditional schedules with identical correction counts at every optimizer step and matched total exposure end with G between 0.1775 and 0.9900; the selected high- and low-transfer schedules replay exactly. Thus, in these tasks, evidence organization changes the held-out mapping, and one fixed-source intervention shows that concrete-minibatch order can also change it under matched macro exposure. The results are behavioral and establish neither an internal mechanism nor a universal or cross-model law.

1 INTRODUCTION

Training accuracy tells us whether a model has matched the observed labels, but not which predictive rule it will apply elsewhere. In an overparameterized model, several functions can agree on every training example and disagree on unseen inputs. We ask what properties of the evidence and training process select among these fitting functions. Throughout the paper, *function* refers only to the model's prespecified A/B mapping on held-out inputs, not to an identified symbolic program or internal representation. Likewise, *fit* refers to the reported training classification accuracies on named groups; matched fit does not mean equal cross-entropy loss, predictive probabilities, or logits.

Contradictory examples provide a useful instrument for this question. The same number of labels can conflict with a primary relation while differing in whether the exceptions share a second relation. We call the aligned fraction *exception coherence* c. Stable one-token identities make instance-specific fitting accessible. The main identity splits test whether the A/B mapping extends beyond observed identity tokens while feature cells recur in training; a separate preregistered control withholds complete joint cells. The construction holds exception quantity and exposure fixed while changing how the contradictory evidence is organized.

The experiments follow this ambiguity through three settings. Study I mixes primary and exception examples and asks whether coherence changes held-out transfer. Family 1 gives a strict five-point response, while Family 2 preserves the positive $c = 1$ versus $c = 0$ endpoint effect but not the graded curve. Study II establishes the primary predictor first and then adds corrections. Although all 21 Phase-2 arms fit their training data, $(N = 200, c = 1)$ transfers perfectly in every seed and $(N = 100, c = 1)$ is highly variable. That instability motivates exploratory, post-hoc fixed-

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.

source follow-ups: path crossover tests whether later training erases earlier history, and the q-profile experiment changes concrete minibatch order while holding the correction count at every optimizer step fixed.

Across these settings, held-out mappings differ despite capacity-qualified near-ceiling fit in Family 1 and exact training classification fit elsewhere. Shared relational structure changes the mapping under the tested interventions, including a complete-cell holdout; one fixed-source micro-order intervention does so under pointwise-matched correction counts. These behavioral results cover one small model and two synthetic task constructions; mechanism and broader generality remain open.

2 Related Work

Several lines of work motivate separating terminal training fit from behavior on held-out inputs.

2.1 Interpolation, Random Labels, And Memorization

Modern networks can interpolate random labels, demonstrating that exact fit alone does not identify a transferable rule (Zhang et al., 2017); double-descent analyses likewise show that interpolation need not imply poor test performance (Belkin et al., 2019). Work separating memorization from pattern learning finds that networks can memorize noise while preferentially learning simpler regularities in real data (Arpit et al., 2017); noisy-label studies also show that fitting order changes over training (Liu et al., 2020). These results motivate evaluating held-out behavior rather than inferring a predictive rule from training accuracy.

2.2 Inductive Bias And Function Selection

An interpolating parameterization can represent many functions, so generalization depends on biases over which function is selected. One account argues that the parameter–function map places greater prior weight on simple functions (Valle-Pérez et al., 2019). This frames selection among fitting functions; our experiments vary a visible property of the evidence rather than characterize the optimization bias responsible for the resulting selection.

2.3 Data-Driven Rule Selection

Rule learning has long been framed as choosing among structured hypotheses: Bayesian concept-learning models infer over compositional rule spaces and predict generalization from the selected rule (Goodman et al., 2008), the Tolerance Principle relates productivity to the amount of exceptional evidence (Yang, 2016), and past-tense acquisition work studies regular and irregular classes (Bybee & Slobin, 1982). A recent small-language-model test finds graded rather than discontinuous behavior near the Tolerance Principle's predicted boundary (Friedman et al., 2026). Study I instead holds exception quantity fixed, while Study II varies quantity and coherence on a coarse prespecified grid rather than estimating a productivity threshold.

2.4 Rules, Facts, And Revision Scope

The Rules-and-Facts model formalizes simultaneous teacher-rule recovery and random-fact memorization (Farné et al., 2026); relational tasks likewise show that generalization and exception memorization can coexist (Cheng & Lippl, 2026). The closest revision-scope account is the v1 *Lifted Representation Hypothesis* (Park & Choi, 2026), which studies whether updates affect shared latent structures rather than isolated facts. Our intervention changes the relational organization of the evidence and measures held-out transfer.

2.5 Rule-Level Editing, Dynamics, And Data Order

RuleEdit evaluates whether edits transfer to rule-derived instances and proposes a dedicated editing method (Zhou et al., 2025); subsequent work studies distributed multi-layer interventions for the same objective (Wang et al., 2026). Our experiments instead use ordinary full-parameter fine-tuning.

Related dynamics work shows memorization and generalization emerging on different schedules (Du et al., 2025; Power et al., 2022); our auxiliary dynamics result is early-only and horizon-sensitive.

Data presentation order is a direct neighbor. Bengio et al. (2009) sequence reweighted training distributions; their Perceptron study orders 200 examples, while their neural studies also change effective support or exposure. Shumailov et al. (2021) show that reordering preformed minibatches can alter learning while preserving batch contents and epoch-level exposure; their reshuffling and replacement attacks relax those controls. Our exploratory follow-up fixes one source, the dataset and concrete-minibatch multiset, per-example exposure, and correction count at every step; all six schedules fit both training groups exactly. This is a conditional comparison, not a priority or general-law claim.

Prior work establishes that neural networks can interpolate training data, memorize exceptions, generalize rules, and favor some fitting functions over others. Our question is narrower and data-level: when fit is attainable and evidence quantity is controlled, does the relational organization and presentation of contradictory examples change the mapping expressed on held-out inputs? We answer this behaviorally, without proposing an editing method or inferring a particular internal representation.

3 Problem Formulation

3.1 Family-1 Task And Data-Generating Functions

Each input is

$$x = (k, p_{\mathrm{L}}, p_{\mathrm{R}}, s_{\mathrm{L}}, s_{\mathrm{R}}), \tag{1}$$

where k is a one-token identity, $p_{\mathrm{L}}, p_{\mathrm{R}} \in \{0, 1\}$ are primary attributes, and $s_{\mathrm{L}}, s_{\mathrm{R}} \in \{1, \ldots, 5\}$ are secondary attributes. The programmed primary function is

$$R_1(x) = \begin{cases} \mathrm{A}, & p_{\mathrm{L}} = p_{\mathrm{R}}, \\ \mathrm{B}, & p_{\mathrm{L}} \neq p_{\mathrm{R}}. \end{cases} \tag{2}$$

The secondary region is selected by

$$z(x) = \mathbb{1}[s_{\mathrm{L}} = s_{\mathrm{R}}], \tag{3}$$

and its conflicting target is $R_2(x) = \mathrm{flip}(R_1(x))$ whenever $z(x) = 1$. These functions define the data and evaluation regions. In particular, R_1 is not evidence that the Study I model has already acquired the primary behavior before seeing exceptions.

3.2 Study I, Family 1: Joint Exception Learning

Study I trains primary-rule and exception examples together. For coherence condition c, let E_c^{I} be an exception set of exactly 200 among 1,000 training rows. The joint supervised target is

$$y_c^{\mathrm{I}}(x) = \begin{cases} \mathrm{flip}(R_1(x)), & x \in E_c^{\mathrm{I}}, \\ R_1(x), & x \notin E_c^{\mathrm{I}}. \end{cases} \tag{4}$$

Coherence $c \in \{0, .25, .5, .75, 1\}$ is the fraction of the fixed exception budget deliberately selected from $z(x) = 1$; the remainder has deterministic random provenance. Random-provenance rows may incidentally satisfy $z(x) = 1$, so provenance and selector overlap remain separately recorded. Every condition keeps exception count, total target balance, prompts, and exposure fixed.

Let $\mathcal{D}_{R_2}^{\mathrm{held}}$ contain secondary-region rows with identity tokens absent from training. At seed s and its terminal t_s, define the binary prediction

$$\widehat{y}_{c,s}(x) = \arg\max_{y \in \{\mathrm{A},\mathrm{B}\}} f_{c,s,t_s}^{\mathrm{I}}(y \mid x), \tag{5}$$

and its held-out accuracy

$$g_s(c) = \frac{1}{|\mathcal{D}_{R_2}^{\mathrm{held}}|} \sum_{x \in \mathcal{D}_{R_2}^{\mathrm{held}}} \mathbb{1}[\widehat{y}_{c,s}(x) = R_2(x)]\,. \tag{6}$$

Thus $g_s(c)$ is simply the binary accuracy on the held-out R_2 set. The original qualification contrast is

$$\Delta_s = g_s(1) - g_s(0). \tag{7}$$

Training-set exception accuracy cannot distinguish individual fitting from transferable acquisition, because both can fit E_c^{I}. Held-out R_2 identities provide the behavioral discriminator; identity tokens are disjoint, but the underlying primary/secondary feature cells occur in training, so $g_s(c)$ excludes exact prompt lookup but does not distinguish a learned equality computation from a mapping over familiar feature cells.

Family 2 retains the same joint-learning distinction but replaces the equality selector with a modular-composition relation. Its input and secondary region are

$$x = (k, p_{\mathrm{L}}, p_{\mathrm{R}}, a, b, d), \quad a, b, d \in \{0, \ldots, 4\}, \qquad z_2(x) = \mathbb{1}[d \equiv (a + b) \bmod 5], \tag{8}$$

with the same primary rule $R_1(x)$ as Eq. (2); on $z_2(x) = 1$ the conflicting target is $\mathrm{flip}(R_1(x))$, and random exceptions satisfy $d \not\equiv (a + b) \bmod 5$. As a Study-I experiment it is separate from the sequential protocol below, which reuses the same fixed-slot generator and secondary relation under a distinct Phase-1/Phase-2 design.

3.3 Study II: Sequential Functional Revision

Study II first trains a seed-specific predictor on R_1 alone, yielding held-out primary and secondary baselines $P_{0,s}$ and $G_{0,s}$. Phase 2 then starts from those exact weights with fresh optimization state and introduces N corrections, of which fraction c is compatible with the same modular relation $d = (a + b) \bmod 5$ used by Family 2. The frozen grid is $N \in \{0, 100, 200\}$ and $c \in \{0, .5, 1\}$ for $N > 0$; $N = 0$ is continued R_1 training.

At the exact Phase-2 terminal, $P_s(N, c)$ is held-out R_1 retention and $G_s(N, c)$ is held-out R_2-consistent transfer. The primary contrasts compare $c = 1$ with $c = 0$ at fixed N, and compare fully coherent evidence across $N = 100$ and 200. This is sequential model-weight revision, unlike Study I's joint-learning curve $g_s(c)$; neither result is pooled with the other.

The training-path follow-ups hold one qualified Phase-1 predictor and one $N = 100, c = 1$ correction problem fixed; they are specified in Section 4 and do not redefine the Study-II endpoint analysis.

4 Experimental Design

4.1 Family-1 Generator And Splits

The balanced base world contains 100 joint primary/secondary attribute cells with ten rows each, 500 A and 500 B labels under the primary rule, and exactly 200 secondary-selector rows. Each seed assigns 1,800 unique one-token identity keys across 1,000 training rows and two 400-row held-out sets; train and test keys are disjoint, but all held-out feature cells occur in training.

The completed generator includes $c \in \{0, .25, .5, .75, 1\}$ ($c = .25$ and $c = .75$ use 50/150 and 150/50 deliberate/random allocations); the $c = 0$ vs $c = 1$ contrast qualifies the endpoint. Every condition has 200 exceptions, 500 A and 500 B targets, identical row order, and row-identical 20-token prompts plus a one-token response. Token identities and exposures are matched, but the one-token vocabulary items may retain pretrained semantics. Generator tests cover identity disjointness, prompt collisions, target balance, exposure, and fixed lengths. A surface-remapping control is not complete.

Feature-combination-OOD control. A separate preregistered control holds out complete joint cells over $c \in \{0, .25, .5, .75, 1\}$ on one fixed 50/50 cell partition, retaining every atomic feature value and all 25 secondary pairs in training; the $c = 0$ versus $c = 1$ contrast is confirmatory and other levels descriptive (Appendix B). Its own frozen capacity ladder qualified all three seeds at step 1,600, so no seed-specific main-experiment horizon is reused.

4.2 Family 2 Intervention

Family 2 is a fixed-slot construction. Each fresh seed has 1,000 training rows and the same 200 identity-defined exception rows at every $c \in \{0, .25, .5, .75, 1\}$; each adjacent level converts 50

Table 1: Frozen optimization and evaluation settings.

Setting	Value
Optimizer	AdamW
Learning rate	2×10^{-5}
Adam betas / epsilon	$(0.9, 0.95)$ / 10^{-8}
Weight decay / gradient clip	0.01 / 1.0
Batch / gradient accumulation	8 / 1
Warmup	20 optimizer steps
Schedule	Linear warmup, then cosine decay
Scheduler horizon	6,400 optimizer steps
Evaluation / checkpoint cadence	100 / 400 steps

nested exception rows by changing only d, while identity, primary features, a, b, labels, row position, and the 800 nonexception rows remain fixed. The audited construction preserves $a/b/d$ marginals, target balance, prompt-token multisets, and token-length distributions, so converted prompts differ at exactly one d-token position; we do not claim control of every latent property. Each seed has fixed 400-row primary- and secondary-unseen sets disjoint in identity from training, with step-3,200 accuracies $P_s(c)$ and $G_s(c)$.

4.3 Model And Optimization

We start each run from the same local Qwen2.5-0.5B base checkpoint (Yang et al., 2024; Qwen Team, 2024). Training is full-parameter BF16 response-only supervised fine-tuning with eager attention and gradient checkpointing. The loss is full-vocabulary causal-LM cross-entropy on the sole response token. Table 1 reports the frozen optimization settings; lower-level determinism and hash fields remain in the archived run manifests.

For each Family-1 seed, we train $c = 0$ first and apply the frozen capacity ladder. The first rung satisfying random-exception memorization and every control sets that seed's endpoint, which all five coherence conditions then share from the immutable base checkpoint with the same scheduler and batch prefix (seeds 101 and 202 at 1,600 steps, seed 303 at 3,200 steps). Data and training randomization vary together across seeds 101, 202, and 303, so the two sources of variation are not separately identified.

Family 2 uses fresh seeds 1001, 1002, and 1003. All 15 arms start from the same immutable base and use the optimization settings in Table 1, but run to the common frozen endpoint of 3,200 steps. There is no capacity ladder, early stop, best-checkpoint selection, or seed-specific horizon in this family.

Study II uses fresh seeds 1101–1103 and reuses the Family-2 fixed-slot modular-composition generator. Phase 1 trains 1,000 all-R_1 examples for exactly 3,200 steps and saves only the qualified seed-matched weights. Phase 2 starts seven arms per seed from those weights with fresh optimizer, scheduler, and RNG state: $N = 0$, and the six combinations of $N \in \{100, 200\}$ with $c \in \{0, .5, 1\}$. Correction identities and labels are fixed across c at each N, the $N = 100$ set is nested within $N = 200$, and every arm uses the same 3,200-step within-seed batch prefix with no endpoint selection.

4.4 Fixed-Source Training-Path Follow-Ups

The large $N = 100, c = 1$ variation motivates two exploratory, post-hoc follow-ups. The first fixes one Phase-1 checkpoint and correction problem and deterministically replays a high-transfer path H and a low-transfer path L. At $t \in \{750, 1000, 1250, 1500\}$, a hybrid follows one path through update t and the opposite path's absolute-step minibatch schedule from update $t + 1$; weights, AdamW moments, scheduler state, global step, and prefix-seeded RNG state continue through the switch.

The second follow-up isolates *micro-order* more tightly. It fixes the same Phase-1 source, $N = 100, c = 1$ dataset, and multiset of 3,200 concrete minibatches across six schedules with identical per-example exposure, pointwise-identical correction count q_t, and fixed model RNG, optimizer, scheduler, and applied learning-rate sequence; only the assignment among equal-q_t positions changes. The highest- and lowest-transfer micro-orders were then selected for exact re-execution.

4.5 Evaluation, Controls, And Formal Analysis

Evaluation scores the A and B candidate tokens at the response position and reports binary accuracy. Family 1 reports $g_s(c)$ on secondary-unseen examples. Family 2 reports secondary transfer $G_s(c)$ and primary retention $P_s(c)$ at exact step 3,200; the within-seed $G_s(1) - G_s(0)$ contrast is the cross-rule-family confirmatory comparison.

The frozen Family-1 validity gates require a 100% generator-test pass rate, at least 0.95 random-exception memorization under $c = 0$, at least 0.90 primary-train and primary-unseen accuracy, at most 0.60 $c = 0$ secondary-unseen leakage, at least 0.95 coherent-train accuracy under $c = 1$, finite numerics, exact repeated evaluations at required rungs, and matching model, tokenizer, schedule, batch-prefix, and horizon records. Conditional on these controls, the signal gates are $c = 1$ secondary-unseen accuracy of at least 0.80 and a $c = 1 - c = 0$ gain of at least 0.20 in every seed. For the OOD control, the fit, primary-OOD (≥ 0.80), lineage, repeat, numerical, and resource gates apply to the six P1 $c = 0$ and $c = 1$ endpoint runs; the intermediate descriptive conditions are gated only on P1 validity plus the same integrity and comparability audits (Appendix B).

For Family 2, all 15 runs must pass data, lineage, schedule, numerical, deterministic-repeat, and resource audits. Every run must fit primary-train and all-exception examples, including every nonempty coherent and random subgroup. The frozen outcome precedence returns a graded replication only under strict aggregate and per-seed monotonicity; otherwise a consistent positive endpoint can support only the endpoint-effect classification. $P(c)$ and all response shape fits are secondary or descriptive and cannot change that verdict.

The Family-1 formal analysis reports every seed endpoint, across-seed means and descriptive Student-t intervals with two degrees of freedom, fixed linear/quadratic summaries, and rank correlations; these describe three seeds rather than provide a high-powered population test.

Study II evaluates 400 fixed primary-unseen and 400 secondary-unseen rows; Phase 1 requires primary-train accuracy at least .99 and $P_0 \geq .95$ in every seed, and Phase 2 requires primary-train and correction-train accuracy at least .95 in every arm. The frozen decision rule jointly checks transfer relative to Phase 1, the fixed-N $c = 0$ arm, and continued training, then requires substantial primary retention.

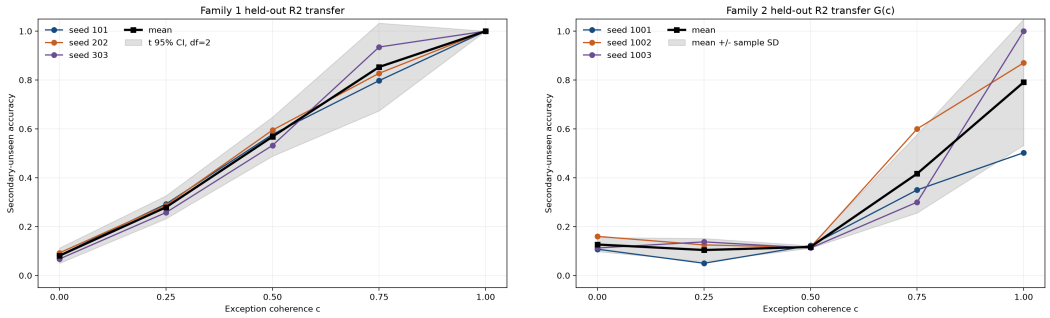
Every crossover hybrid must fit all 900 primary and 100 correction training rows and retain $P \geq .80$; the micro-order analysis applies the same gates to all six schedules and measures terminal G and P at step 3,200. Its frozen descriptive rule labels the effect "strong" only if the terminal-G range is at least 160/400 and the middle-four span after removing one minimum and maximum is at least 81/400; an effect label requires a range of at least 80/400. The selected-extreme replay requires literal equality, with no numeric tolerance, for all 33 G/P measurements, terminal counts, and repeated prediction digests. Neither follow-up recomputes the Study-II analysis.

5 Results

5.1 Layer I: Shared Structure At Fixed Evidence Quantity

Family 1. The $c = 0$ condition first establishes that the backbone and optimization can fit the tested deterministic random-provenance exceptions: random-provenance and pure-random training accuracy are both 1.000 in every seed, while held-out R_2 transfer is only 0.0825, 0.0925, and 0.0675 (Appendix Tables 4–5), so low transfer does not coincide with failure to fit. When all 200 exceptions instead align with R_2, coherent-train and held-out transfer both reach 1.000 in every seed, with within-seed gains of 0.9175, 0.9075, and 0.9325. Across all five coherence levels the means are 0.0808, 0.2792, 0.5683, 0.8533, and 1.0000 (Figure 1), strictly increasing in every seed and with each condition sharing its seed's selected horizon. Appendix Table 4 shows that the full comparison has near-ceiling, not uniformly exact, fit: primary-nonexception accuracy is .99625–1.000, all-exception accuracy is .980–1.000, and 9 of 15 runs classify all rows correctly.

Feature-combination-OOD control. In the preregistered complete-cell holdout, secondary-OOD accuracy at $c = 0$ versus $c = 1$ is .1175 versus 1.0000, .0875 versus 1.0000, and .0775 versus .9825 (means .0942 versus .9942; mean gain .9000); all six runs pass the fit and primary-OOD gates. The descriptive five-level response is strictly increasing (Appendix Table 7).



(a) Family 1 (b) Family 2

Figure 1: Study-I secondary transfer. Family 1 shows a strict five-point response at each seed's frozen terminal; Family 2 (step 3,200) retains the observed low-coherence reversals. Lines show seeds and mean; bands are a descriptive Student-t 95% interval in (a) and mean $\pm$ sample SD in (b).

Family 2 (modular composition). Family 2 tests $d = (a + b) \bmod 5$ with three fresh seeds while keeping exception identities and labels fixed across coherence conditions. Adjacent levels convert 50 of the same 200 exception rows by changing only d.

At step 3,200, every seed has higher G at $c = 1$ than at $c = 0$, with gains 0.3950, 0.7100, and 0.8875 (Appendix Table 6). Mean $G(c)$ is 0.1267, 0.1042, 0.1167, 0.4167, and 0.7908. Neither the mean nor any seed is strictly increasing over all five levels. The frozen classification is `CROSS_FAMILY_ENDPOINT_EFFECT_ONLY`: every seed shows a positive $c = 1 - c = 0$ contrast, but the strict graded-replication criterion fails. All 15 arms fit every defined training subgroup to 1.000; across the two families, shared structure changes transfer scope. The response shapes differ between the two task constructions, which also operationalize coherence differently; these experiments do not isolate the source of that difference.

5.2 Layer II: Sequential Functional Revision

All three Phase-1 predictors establish R_1 at step 3,200, with primary-train and held-out $P_0 = 1.000$ and secondary $G_0 = 0.000$. All 21 Phase-2 arms then fit primary training rows and every nonempty correction set to 1.000, so their held-out differences occur under equal terminal training fit; the complete matrix appears in Appendix A.

At $N = 200, c = 1$, every seed reaches $P = G = 1.000$. At $N = 200, c = 0$, mean $(P, G) = (.8983, .1083)$, giving planned mean contrasts of $\Delta P = +.1017$ and $\Delta G = +.8917$ (Figure 2). The smaller coherent intervention is a material boundary: at $N = 100, c = 1$, per-seed G is 0.0125, 0.7825, and 0.0300 (mean .2750, sample SD .4396). The two tested positive quantities do not identify a correction threshold, and Nc is not established as a sufficient statistic.

5.3 Layer III: Training Path And Micro-Order

The $N = 100, c = 1$ instability motivates exploratory, post-hoc training-path tests. For one fixed Phase-1 source and correction problem, a high-transfer path H ends at $(G, P) = (.9700, 1.0000)$ and a low-transfer path L at $(.0325, .9825)$; deterministic replay reproduces both exactly. Crossing their future schedules at steps 750, 1,000, 1,250, and 1,500 yields H$\rightarrow$L terminal G of .6350, .9650, .9250, and .9850, whereas L$\rightarrow$H yields .0225, .0300, .0250, and .0450 (Figure 3; Appendix Table 8). Every hybrid fits both training groups to 1.000 and retains $P = .9725$–1.0000, so early path history strongly constrains the later held-out mapping here.

The q-profile intervention isolates presentation order more tightly. Six fixed schedules share the same source, dataset, 3,200 concrete minibatches, per-example exposure, model RNG, optimizer, scheduler, and applied learning rates, with the same correction count q_t at every step; only the assignment among equal-q_t positions changes. In fixed schedule order, terminal G is .6125, .7475, .7025, .5525, .9900, and .1775, while P is .9975, .9975, .9975, .9975, 1.0000, and .9500; all six fit all 900 primary and 100 correction rows to 1.000.

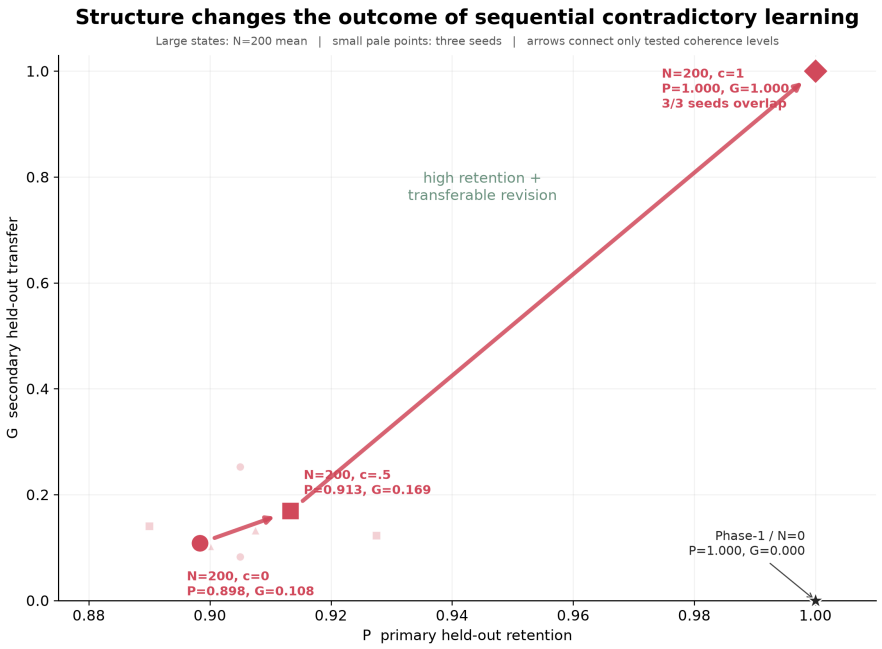


Figure 2: Study-II sequential revision at $N = 200$ (step 3,200): P is held-out R_1 retention, G held-out R_2 transfer. Pale markers are seed endpoints, large markers their means at $c \in \{0, .5, 1\}$. The mean moves from $(P, G) = (.898, .108)$ at $c = 0$ to $(1, 1)$ at $c = 1$; arrows connect tested means only and imply neither interpolation nor a threshold.

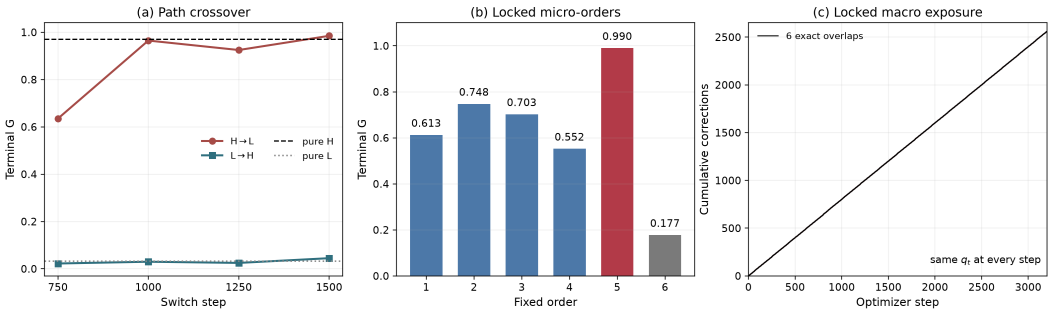


Figure 3: Exploratory fixed-source training-path evidence. (a) Crossover terminal G; dashed lines mark replayed H/L references. (b) Terminal G for six fixed micro-orders; all schedules fit both training groups to 1.000. (c) Exactly overlapping cumulative correction exposure (same q_t every step). Same macro correction timing, different micro-order, different held-out mapping; the six schedules are conditional interventions, not population replications.

The six orders span G from .1775 to .9900. The prespecified descriptive effect criterion is met, though the stronger anti-concentration criterion is not; this does not imply absence of an order effect (Section 4; Appendix E). Exact re-executions of the selected high- and low-transfer extreme orders reproduce all 33 G/P measurements, terminal counts, and repeated prediction digests literally exactly. This is controlled behavioral evidence that concrete-minibatch ordering can change the terminal held-out mapping in this fixed setting under the locked macro exposure profile.

6 Discussion

Prior work establishes the broad fit–generalization dissociation. Our narrower contrasts hold exception quantity and exposure fixed: relational organization changes held-out transfer under near-ceiling

Family-1 and exact Family-2 fit. In Study II, exact named-group fit coexists with held-out G from near zero to 1.000. A preregistered complete-cell holdout reproduces the Family-1 endpoint contrast.

Family 2 reproduces the positive endpoint contrast, but not Family 1's strictly increasing response. The task constructions also operationalize coherence differently, so the source of their shape difference is not isolated.

The micro-order experiment fixes the source, dataset, optimizer configuration, applied learning-rate sequence, concrete-minibatch multiset, per-example exposure, and correction count at every step; reassigning minibatches among equal-count positions still changes the terminal mapping. Exact replay confirms the selected schedules, but not a mechanism or population law. The coarse Study-II grid leaves the 100–200-correction interval unresolved.

7 LIMITATIONS

- **Scope and data.** All experiments use Qwen2.5-0.5B, synthetic binary tasks, two rule families, and one optimizer. No cross-scale, cross-model-family, additional-rule-family, natural-language corrective-SFT, instruction-following, or open-generation replication is complete.
- **Holdout and interface.** The main identity holdouts reuse feature cells. The preregistered OOD control withholds complete joint cells but uses one fixed 50/50 partition, ten held diagonal cells, and a descriptive five-level curve. Evaluation is restricted to an A/B choice, and no surface-remap control is complete.
- **Interpretation.** The evidence concerns held-out behavior. It identifies neither a symbolic rule nor a latent representation, reusable interface, or neural/gradient mechanism. Coherence is a known generator parameter, operationalized differently by the two families, rather than a universal metric.
- **Sequential and path scope.** Study II uses a coarse $N \times c$ grid, which identifies no correction threshold and does not establish Nc as a sufficient statistic or scaling law. The crossover and micro-order studies fix one Phase-1 source and one $N = 100, c = 1$ problem; the six orders are conditional interventions, the replayed extremes were selected after observing their spread, and concrete batch properties beyond q_t move with batch identity.

8 CONCLUSION

Beyond the established fit–generalization dissociation, this paper contributes two controlled contrasts. At fixed exception quantity and exposure, relational organization changes held-out A/B transfer under near-ceiling or exact training classification fit. In one exploratory fixed-source intervention, matched minibatch multiset, per-example exposure, pointwise correction count, and exact terminal accuracy still permit different mappings across micro-orders. A preregistered complete-cell holdout extends the Family-1 endpoint contrast; broader scope and mechanism remain open.

AI Use Statement. Generative AI tools substantially assisted hypothesis and control design; implementation and debugging of experimental, synthetic-data, and analysis code; literature discovery and summarization; interpretation; and manuscript drafting and editing. Reported numerical results were computed from recorded training and evaluation outputs using analysis scripts. AI-assisted outputs were checked against experimental artifacts and recorded results. The authors take responsibility for the final methodology, claims, results, and manuscript.

Reproducibility Statement. Fixed synthetic generators, seeds, schedules, endpoints, and validity gates support reproducibility; the appendix reports principal endpoint matrices and controls. Machine-readable configurations, results, and code are retained in the project artifacts.

REFERENCES

Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 233–242. PMLR, 2017. URL https://proceedings.mlr.press/v70/arpit17a.html.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019. doi: 10.1073/pnas.1903070116. URL https://doi.org/10.1073/pnas.1903070116.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 41–48. Association for Computing Machinery, 2009. doi: 10.1145/1553374.1553380. URL https://doi.org/10.1145/1553374.1553380.

Joan L. Bybee and Dan I. Slobin. Rules and schemas in the development and use of the english past tense. *Language*, 58(2):265–289, 1982. doi: 10.2307/414099. URL https://doi.org/10.2307/414099.

Luke Cheng and Samuel Lippl. A mathematical theory of balancing relational generalization and memorization, 2026. URL https://arxiv.org/abs/2605.22972v1.

Yupei Du, Philipp Mondorf, Silvia Casola, Yuekun Yao, Robert Litschko, and Barbara Plank. Reason to rote: Rethinking memorization in reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 8659–8679, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.437. URL https://aclanthology.org/2025.emnlp-main.437/.

Gabriele Farné, Fabrizio Boncoraglio, and Lenka Zdeborová. The rules-and-facts model for simultaneous generalization and memorization in neural networks, 2026. URL https://arxiv.org/abs/2603.25579v1.

Adam E. Friedman, Stevan Harnad, and Rushen Shi. Tolerance principle and small language model learning. In Romi Hill, Xinhan Jiang, Danutham Worapipat, and Aditya Yedetore (eds.), *Proceedings of the 50th Annual Boston University Conference on Language Development*, pp. 234–247, Somerville, MA, 2026. Cascadilla Press. URL https://www.lingref.com/bucld/50/BUCLD50-19.pdf.

Noah D. Goodman, Joshua B. Tenenbaum, Jacob Feldman, and Thomas L. Griffiths. A rational analysis of rule-based concept learning. *Cognitive Science*, 32(1):108–154, 2008. doi: 10.1080/03640210701802071. URL https://doi.org/10.1080/03640210701802071.

Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. In *Advances in Neural Information Processing Systems*, volume 33, pp. 20331–20342. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/ea89621bee7c88b2c5be6681c8ef4906-Abstract.html.

Bumjin Park and Jaesik Choi. Lifted representation hypothesis in language models, 2026. URL https://arxiv.org/abs/2607.19360v1. arXiv:2607.19360v1; ICML 2026 Workshop: Philosophy Meets Machine Learning.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets, 2022. URL https://arxiv.org/abs/2201.02177.

Qwen Team. Qwen2.5-0.5B. Hugging Face model card, 2024. URL https://huggingface.co/Qwen/Qwen2.5-0.5B. Base pretrained checkpoint; accessed 2026-08-15.

Ilia Shumailov, Zakhar Shumaylov, Dmitry Kazhdan, Yiren Zhao, Nicolas Papernot, Murat A. Erdogdu, and Ross J. Anderson. Manipulating SGD with data ordering attacks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18021–18032. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/959ab9a0695c467e7caf75431a872e5c-Abstract.html.

Guillermo Valle-Pérez, Chico Q. Camargo, and Ard A. Louis. Deep learning generalizes because the parameter-function map is biased towards simple functions. In *International Conference on Learning Representations*, 2019. doi: 10.48550/arXiv.1805.08522. URL https://openreview.net/forum?id=rye4g3AqFm.

Yating Wang, Wenting Zhao, Yaqi Zhao, Yongshun Gong, Yilong Yin, and Haoliang Sun. Distributed multi-layer editing for rule-level knowledge in large language models, 2026. URL https://arxiv.org/abs/2604.08284. Under review.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. doi: 10.48550/arXiv.2412.15115. URL https://arxiv.org/abs/2412.15115.

Charles Yang. *The Price of Linguistic Productivity: How Children Learn to Break the Rules of Language*. The MIT Press, Cambridge, MA, 2016. doi: 10.7551/mitpress/9780262035323.001.0001. URL https://doi.org/10.7551/mitpress/9780262035323.001.0001.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017. doi: 10.48550/arXiv.1611.03530. URL https://openreview.net/forum?id=Sy8gdB9xx.

Bihan Zhou, HaoPeng Ren, Li Yuan, Yi Cai, Liuwen Cao, and Zikun Deng. RuleEdit: Towards rule-level knowledge generalization to mitigate over-editing in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 3159–3175, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.164. URL https://aclanthology.org/2025.findings-acl.164/.

A Study II Endpoint Matrix

Study II uses three fresh seed-specific Phase-1 predictors. At exact step 3,200, each has primary-train accuracy and held-out $P_0 = 1.000$, with $G_0 = 0.000$. Table 2 reports every Phase-2 endpoint; no arm, seed, or checkpoint was selected. Every listed arm has primary-train accuracy 1.000, and every nonempty correction set has correction-train accuracy 1.000.

Table 2: Complete Study-II Phase-2 endpoints at exact step 3,200. P is held-out primary retention and G is held-out secondary transfer.

Arm	N	c	P_{1101}	P_{1102}	P_{1103}	G_{1101}	G_{1102}	G_{1103}
Continued R_1	0	–	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000
Random-100	100	0	0.9750	0.9900	0.9800	0.0300	0.0225	0.0150
Mixed-100	100	.5	0.9825	0.9775	0.9625	0.0300	0.0225	0.0350
Coherent-100	100	1	0.9950	1.0000	0.9775	0.0125	0.7825	0.0300
Random-200	200	0	0.9050	0.8900	0.9000	0.0825	0.1400	0.1025
Mixed-200	200	.5	0.9050	0.9275	0.9075	0.2525	0.1225	0.1325
Coherent-200	200	1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

The frozen planned contrasts are shown in Table 3. The primary parameterization is (N, c); the derived coherent count $K = Nc$ does not define a universal law. Exact step 3,200 is the sole confirmatory endpoint.

Table 3: Frozen Study-II planned contrasts, reported as three-seed mean difference $\pm$ sample SD.

Contrast	ΔP	ΔG
Coherence at $N = 100$: $(100, 1) - (100, 0)$	0.0092 ± 0.0113	0.2525 ± 0.4398
Coherence at $N = 200$: $(200, 1) - (200, 0)$	0.1017 ± 0.0076	0.8917 ± 0.0292
Coherent quantity: $(200, 1) - (100, 1)$	0.0092 ± 0.0118	0.7250 ± 0.4396
Random load: $(200, 0) - (100, 0)$	-0.0833 ± 0.0153	0.0858 ± 0.0325
Same $K = 100$: $(200, .5) - (100, 1)$	-0.0775 ± 0.0109	-0.1058 ± 0.4848

B Complete Study-I, Family 2, And OOD Tables

Table 4 reports realized Family-1 training classification fit at every coherence level. It is not uniformly exact: 9 of 15 runs classify all 1,000 training rows correctly. Across the 15 runs, primary-nonexception accuracy ranges from .99625 to 1.000 and all-exception accuracy from .980 to 1.000. Among nonempty subgroups, coherent, random-provenance, pure-random, and random- overlap accuracy have minima .9933, .9400, .9250, and 1.0000, respectively. These are classification accuracies, not claims of equal cross-entropy loss, probabilities, or logits.

Table 4: Family-1 terminal training-fit audit. Entries are correct/total A/B classifications at each seed's frozen terminal (steps 1,600, 1,600, and 3,200 for seeds 101, 202, and 303). "Primary" denotes the 800 nonexception rows and "Exception" all 200 exception rows.

Seed	Group	$c = 0$	$c = .25$	$c = .50$	$c = .75$	$c = 1$
101	Primary	799/800	800/800	800/800	800/800	800/800
	Exception	200/200	200/200	199/200	196/200	200/200
202	Primary	800/800	800/800	797/800	799/800	800/800
	Exception	200/200	199/200	199/200	197/200	200/200
303	Primary	800/800	800/800	800/800	800/800	800/800
	Exception	200/200	200/200	200/200	200/200	200/200

Table 5: Secondary-unseen restricted A/B accuracy over the complete Family-1 coherence intervention.

Seed	$c = 0$	$c = .25$	$c = .50$	$c = .75$	$c = 1$
101	0.0825	0.2925	0.5775	0.7975	1.0000
202	0.0925	0.2875	0.5950	0.8275	1.0000
303	0.0675	0.2575	0.5325	0.9350	1.0000
Mean	0.0808	0.2792	0.5683	0.8533	1.0000

Table 6: Family 2 endpoints at the frozen step 3,200. G is secondary-unseen transfer and P is primary-unseen retention; both are restricted A/B accuracies.

Outcome	Seed	$c = 0$	$c = .25$	$c = .50$	$c = .75$	$c = 1$
G	1001	0.1075	0.0500	0.1225	0.3500	0.5025
G	1002	0.1600	0.1250	0.1150	0.6000	0.8700
G	1003	0.1125	0.1375	0.1125	0.3000	1.0000
G	Mean	0.1267	0.1042	0.1167	0.4167	0.7908
P	1001	0.8775	0.9475	0.8800	0.9625	0.9700
P	1002	0.8575	0.8625	0.9125	0.9675	0.9975
P	1003	0.9075	0.8725	0.9200	0.8300	1.0000
P	Mean	0.8808	0.8942	0.9042	0.9200	0.9892

Feature-combination-OOD control. The preregistered control withholds complete four-way joint cells while retaining every atomic feature value and all 25 secondary pairs in training. Table 7 reports the full five-level endpoint comparison; the $c = 0$ versus $c = 1$ contrast is the confirmatory result, while the intermediate levels are descriptive and create no new formal verdict. The fit and primary-OOD (≥ 0.80) gates govern the six P1 $c = 0$/$c = 1$ endpoint runs; P2 intermediate conditions were gated on P1 validity and the integrity/comparability audits rather than on their own threshold. Their mean primary-OOD accuracy is .9100, .8642, .9683, .9867, and 1.0000 (e.g., .8250 and .8625 at $c = .25$ for seeds 101 and 202), and nonexception fit ranges from .9900 to 1.0000 across the 15 runs. Figure 4 shows the secondary-OOD curves.

Table 7: Preregistered Family-1 feature-combination-OOD endpoints at step 1,600. Training uses 50 complete joint cells and evaluation the complementary 50; each entry is restricted A/B accuracy over 400 identity-disjoint rows. "Sec." is the held-out secondary split and "Prim." the held-out primary split. The $c = 0$ versus $c = 1$ contrast is confirmatory; the intermediate levels are descriptive. The control has its own frozen capacity ladder, and all seeds qualified at step 1,600, so no seed-specific main-experiment horizon is reused.

Seed	Split	$c = 0$	$c = .25$	$c = .50$	$c = .75$	$c = 1$
101	Sec.	.1175	.4025	.5175	.8925	1.0000
202	Sec.	.0875	.2225	.5125	.8900	1.0000
303	Sec.	.0775	.3250	.6175	.8575	.9825
Mean	Sec.	.0942	.3167	.5492	.8800	.9942
101	Prim.	.9075	.8250	.9875	.9800	1.0000
202	Prim.	.9200	.8625	.9575	.9950	1.0000
303	Prim.	.9025	.9050	.9600	.9850	1.0000
Mean	Prim.	.9100	.8642	.9683	.9867	1.0000

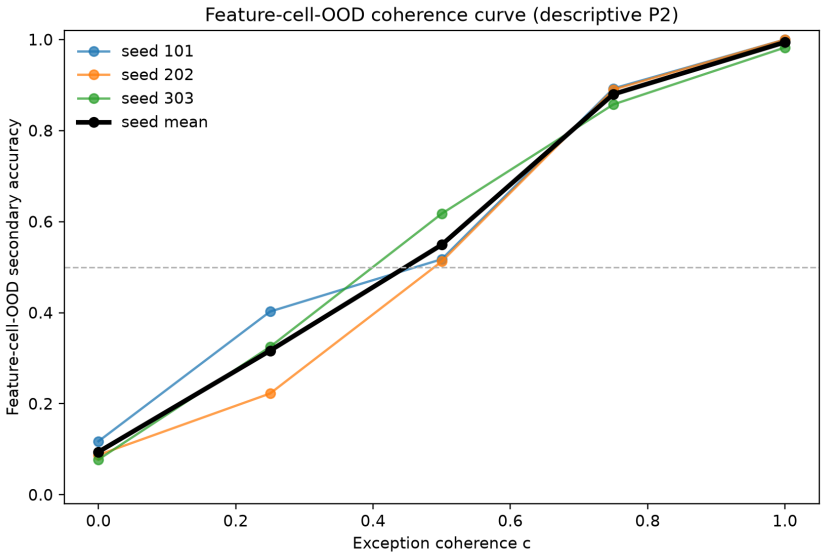


Figure 4: Family-1 feature-combination-OOD secondary transfer over the complete coherence intervention. Lines are the three fixed seeds and the black line their mean; the endpoint contrast is confirmatory and the intermediate levels are descriptive.

C Generator Specifications And Prompt Serialization

Both families serialize a sample through one fixed newline template with one dynamic value per token; the response is a single space-prefixed token. Family 1 uses the template

```
Nonce classification record
Instance: key
Primary: pw[pL] pw[pR]
Secondary: sw[sL] sw[sR]
Answer:
```

where `pw` and `sw` index the frozen one-token primary and secondary symbol banks. A realized secondary-region row is

```
Nonce classification record
Instance: sus
Primary: t t
Secondary: c c
Answer:
```

which is 20 prompt tokens; with the response the sequence is 21 tokens.

Family 2 uses the same frame with a third secondary field,

```
Nonce classification record
Instance: key
Primary: pw[pL] pw[pR]
Secondary: sw[a] sw[b] sw[d]
Answer:
```

and a realized coherent exception row ($a = 3$, $b = 3$, $d = 1$, $R_1 = \mathrm{A}$, label B) is

```
Nonce classification record
Instance: f
Primary: t t
Secondary: the the c
Answer:
```

which is 21 prompt tokens; with the response the sequence is 22 tokens. Within a family, prompt length and token exposure are fixed by construction.

D Additional Descriptive Results

Family 2 response shape. The frozen simple-function comparison ranks logistic, piecewise-linear, then linear. Centered seed-curve correlations are 0.9896, 0.9003, and 0.9093 (median 0.9093), consistent with similar delayed shapes across three seeds. These are descriptive summaries over five sampled levels; they establish neither a logistic law, continuity, nor a threshold. Mean $P(c)$ is 0.8808, 0.8942, 0.9042, 0.9200, and 0.9892, but individual curves do not support a common response function.

Study-I dynamics. Define $D(c, s, t)$ as held-out R_2 accuracy minus pure-random training accuracy. At $c = .5$, its normalized signed area over $[0, 1600]$ is 0.2930, 0.1679, and 0.2816. The late $[1200, 1600]$ values are -0.2534, -0.3863, and -0.3750, and seed 303's full $[0, 3200]$ value is -0.0789. This is an early-only, horizon-sensitive observation, not a stable learning-order law.

E TRAINING-PATH FOLLOW-UP DETAILS

High/low replay and crossover. The fixed high- and low-transfer references have terminal $(G, P) = (.9700, 1.0000)$ and $(.0325, .9825)$. Deterministic replay reproduces both terminal endpoints exactly, with 33-point G and P trajectory correlations of 1.000. Table 8 reports all eight crossover endpoints. Complete unsmoothed trajectories appear in Figures 5 and 6. These exploratory comparisons support temporal path dependence for the tested source, problem, paths, and switch times only.

Table 8: Fixed-source $N = 100, c = 1$ crossover endpoints at step 3,200. Every hybrid has primary- and correction-train accuracy 1.0000 and passes the frozen controls.

Switch	H→L		L→H	
t	G	P	G	P
750	.6350	.9900	.0225	.9875
1,000	.9650	1.0000	.0300	.9800
1,250	.9250	1.0000	.0250	.9725
1,500	.9850	1.0000	.0450	.9750

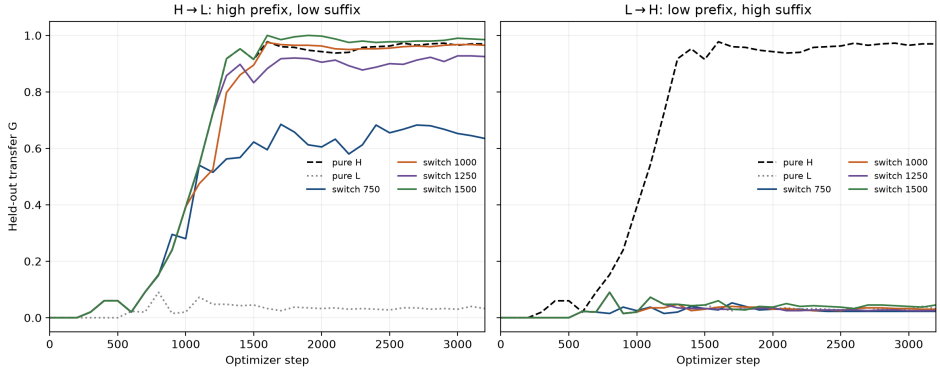


Figure 5: Complete unsmoothed 33-point held-out $G(t)$ trajectories for the replay-confirmed pure H and L paths and all eight fixed crossover hybrids. The left panel shows H→L and the right L→H. These trajectories are follow-up diagnostics and do not alter the formal Study-II endpoint analysis.

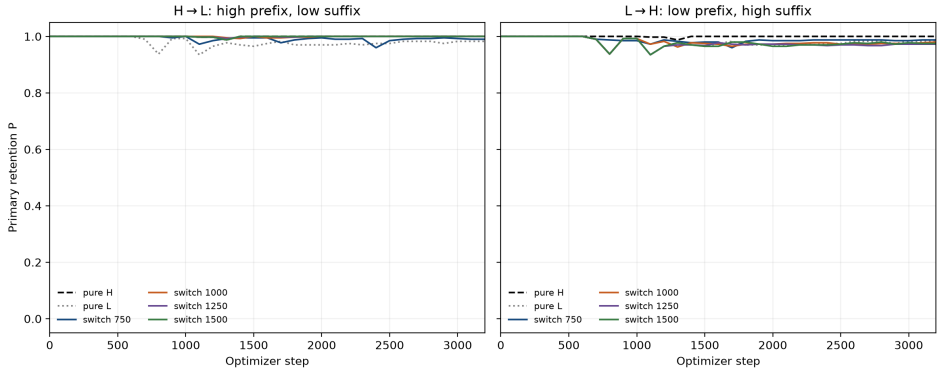


Figure 6: Complete unsmoothed 33-point held-out primary-retention $P(t)$ trajectories for the replay-confirmed references and all eight crossover hybrids. Every hybrid ends with $P \geq .9725$.

Q-profile-locked micro-order. All six schedules contain the same 3,200 concrete minibatches and have identical q_t and cumulative correction exposure at every step, identical per-example total exposure, and the same source, model RNG, optimizer, scheduler, and applied learning-rate sequence. The common q_t histogram has 1,355, 1,273, 449, 105, 17, and 1 steps with $q_t = 0, 1, 2, 3, 4, 5$, respectively; total correction exposure is 2,559. Table 9 reports every endpoint.

Table 9: Exploratory, fixed-source q-profile-locked micro-order endpoints at step 3,200. The six conditional schedules share pointwise q_t, cumulative correction exposure, and the controls described in Section 4. Every order has primary-train accuracy 1.0000 and correction-train accuracy 1.0000.

Order	G	G count	P	P count
1	.6125	245/400	.9975	399/400
2	.7475	299/400	.9975	399/400
3	.7025	281/400	.9975	399/400
4	.5525	221/400	.9975	399/400
5	.9900	396/400	1.0000	400/400
6	.1775	71/400	.9500	380/400

The total terminal-G range is $325/400 = .8125$. Removing one minimum and one maximum leaves counts 221, 245, 281, and 299, whose span is $78/400 = .1950$. Because the frozen strong criterion requires at least $81/400 = .2025$, the exact descriptive classification is `Q_PROFILE_LOCKED_MICRO_ORDER_EFFECT`, not STRONG. This anti-concentration boundary does not erase the observed order effect. The six orders are conditional interventions on one source rather than population replications, and concrete batch properties beyond q_t move with batch identity. Figure 7 shows all unsmoothed G/P measurements.

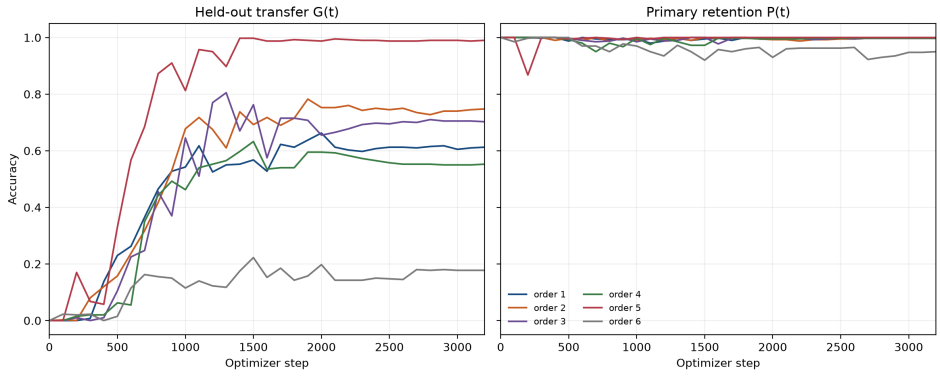


Figure 7: Complete unsmoothed 33-point trajectories for the six q-profile- locked micro-orders. The left panel shows held-out transfer $G(t)$ and the right primary retention $P(t)$. Order labels follow the fixed table order; no smoothing or path selection is applied.

Selected-extreme exact replay. Orders 5 and 6 were selected after the six-order analysis as the observed high and low extremes. Their re-executions reproduce literally exactly all 33 G points, all 33 P points, terminal G/P/primary-train/correction-train counts, and repeated-evaluation prediction-digest payloads at steps 0, 400, 800, 1,600, and 3,200. The applied learning-rate sequence, RNG digests, complete training trace, evaluation tree, final checkpoint contents, and final optimizer/scheduler state also match exactly, with no numeric tolerance. Both replays receive the classification `Q_PROFILE_EXTREME_REPLAY_EXACT`. This verifies reproducible selection by two specific microscopic orders; it is not a new unbiased effect-size sample.