WHEN DOES FINITE EVIDENCE SUPPORT LENGTH GENERALIZATION?

Anonymous authors
Paper under double-blind review

ABSTRACT

Finite evidence is sufficient only relative to a declared deployment protocol. We separate four contracts: whether allowed evidence can identify every deployment-bad alternative, whether the data process acquires decisive evidence, whether the learner recovers an evidence-supported computation, and how remaining discrepancies behave under joint execution. A prefix-trigger family shows that information demand can be non-monotone for evaluation at one horizon but monotone when deployment requires correctness at every horizon up to a limit; in the nontrivial regime, teaching, adaptive querying, and passive sampling cost $1 : d : \Theta(2^d \log(1/\delta))$. A pair-state operator exactly evaluates endpoint disagreement, while its killed restriction measures the stronger never-error event. Their combination yields an end-to-end certificate that adds finite-evidence uncertainty to execution discrepancy. Controlled Recovery diagnostics and neural experiments instantiate these interfaces.

1 INTRODUCTION

Two programs can agree on every short example and still differ after a rare trigger. A teacher who knows the trigger can reveal it once, an adaptive learner can locate it through chosen questions, and a passive process may wait exponentially long.

Whether that distinction matters depends on deployment. If evaluation inspects only one very late endpoint, two trigger systems can look alike again after both outputs saturate. If deployment requires correctness at every horizon up to that point, the earlier distinguishable regime cannot be forgotten. Finite evidence is therefore not sufficient or insufficient in isolation; sufficiency is relative to the horizons and input distributions on which behavior must agree.

One endpoint score also merges failures occurring at different points between evidence and deployment. Decisive evidence may exist but never be acquired. It may be acquired but fitted poorly. Even a nearly correct local computation can fail when its few errors lie on important deployment paths. We organize these contracts as

$$\boxed{\text{Identification} \longrightarrow \text{Acquisition} \longrightarrow \text{Recovery} \longrightarrow \text{Composition.}}$$

Identification asks which computations the allowed channel can distinguish under the deployment protocol. Acquisition asks whether sampling or querying obtains those distinctions. Recovery calibrates how evidence constraints become a learned computation. Composition evaluates that computation along deployment trajectories. These stages can interact; they are neither independent physical causes nor separate parameter modules.

Our theory makes the two ends of this chain share one deployment object. For a $3d$-state prefix-trigger family, point-horizon distinguishability rises and then falls, whereas all-up-to-horizon distinguishability can only grow. At execution time, a pair-state process exactly computes the same endpoint disagreement used by Identification, while a killed operator computes never-error survival. An interface theorem then combines the uncertainty left by finite evidence with the learned computation's joint-execution discrepancy.

The experiments instantiate these contracts. Family-aware Transformers separate sufficient from insufficient prefix-trigger supports, while late-horizon accuracy can rise through saturation without

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

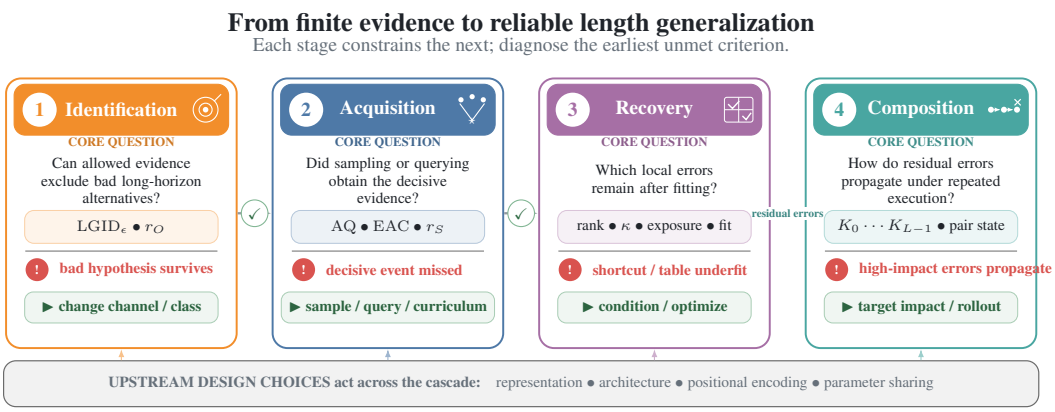


Figure 1: The deployment-conditioned diagnostic cascade. Each contract supplies the object examined by the next; the earliest unmet criterion identifies the first resource that must be addressed.

secret Recovery. HMC models with equal average local Recovery exhibit different finite-horizon survival, and stage-targeted perturbations with similar endpoint scores require different interventions.

Our contributions are:

- **Deployment-conditioned information.** Nested protocols induce monotone certificate and acquisition requirements, while a generalized prefix-trigger theorem separates point-horizon and all-up-to-horizon evaluation.
- **Joint-execution theory.** Pair-state dynamics evaluate endpoint disagreement; killed dynamics, path sensitivity, and repair complementarity characterize the stronger never-error objective.
- **End-to-end interface.** A deployment certificate adds the version-space radius left after Acquisition to the execution discrepancy of a recovered reference computation.
- **Recovery diagnostics and validation.** Controlled rank, conditioning, exposure, and shared-parameter specializations support neural experiments that separate the four contracts.

2 Related Work

Teaching dimension studies the smallest labeled set specifying a target (Goldman & Kearns, 1995); our certificates are channel-restricted, approximate, and indexed by deployment behavior. Passive Acquisition uses hitting and VC ϵ-net ideas (Haussler & Welzl, 1987), while membership queries and observation tables are classical in active automata learning (Angluin, 1987; Isberner et al., 2015). Length-generalization theory studies training lengths, recurrence, positional encodings, representations, and inductive biases that support long-input behavior (Huang et al., 2025; Chen et al., 2025; Izzo et al., 2025; Yang et al., 2026; Weiss et al., 2021; Zhou et al., 2024a; Strobl et al., 2024; Kazemnejad et al., 2023; Zhou et al., 2024b; Fan et al., 2024; Nye et al., 2021; Hou et al., 2025). We instead organize the question around a declared deployment protocol: which distinctions finite evidence must resolve, whether data resolve them, whether learning recovers a corresponding computation, and how its discrepancies translate into reliability.

3 Theory: Four Deployment-Conditioned Contracts

Let $\mathcal{T}$ be a nonempty set of deployment horizons, with input distribution $\mathcal{D}_\ell$ at each $\ell \in \mathcal{T}$. Define

$$d_{\mathcal{T}}(h, g) = \sup_{\ell \in \mathcal{T}} \Pr_{x \sim \mathcal{D}_\ell} [h(x) \neq g(x)].$$

This is a pseudometric, and a metric after identifying computations that agree almost surely at every declared horizon. Evidence S leaves

$$\mathcal{V}_O(S) = \{h : O_e(h) = O_e(f^\star)\ \forall e \in S\}, \qquad r_{O,\mathcal{T}}(S) = \sup_{h \in \mathcal{V}_O(S)} d_{\mathcal{T}}(h, f^\star).$$

For a fixed protocol, we abbreviate the complete-channel and acquired-evidence radii by $r_O = r_{O,\mathcal{T}}(\mathcal{E}^O_{\mathrm{train}})$ and $r_S = r_{O,\mathcal{T}}(S_{\mathrm{acq}})$.

3.1 Identification and Acquisition

What must evidence distinguish? Identification asks whether every computation consistent with the evidence is safe under the entire protocol:

$$\mathrm{LGID}^{O,\mathcal{T}}_\epsilon(f^\star) = \min_S\{|S| : r_{O,\mathcal{T}}(S) \le \epsilon\}.$$

For passive samples $S_m \sim \mu^m$, define $\mathrm{EAC}^{\mathrm{pass},\mathcal{T}}_{\epsilon,\delta}$ by requiring this radius to be at most ϵ with probability $1-\delta$; the target-independent form chooses μ before the target. Active querying instead stops when $\mathrm{diam}_{\mathcal{T}}(\mathcal{V}) = \sup_{h,g\in\mathcal{V}} d_{\mathcal{T}}(h,g) \le \epsilon$; let $\mathrm{AQ}^{O,\mathcal{T}}_\epsilon$ denote its minimum worst-case query count.

Proposition 3.1 (Protocol order and target uniformity)**.** *If $\mathcal{T}_1 \subseteq \mathcal{T}_2$, then $d_{\mathcal{T}_1} \le d_{\mathcal{T}_2}$. Consequently LGID, passive EAC under the same sampling guarantee, and diameter-based AQ cannot decrease when the protocol is enlarged. Moreover, for every version space containing $f^\star$,*

$$r_{f^\star}(\mathcal{V}) \le \mathrm{diam}(\mathcal{V}) \le 2r_{f^\star}(\mathcal{V}).$$

Thus the minimum certificate costs obey

$$C_{\mathrm{rad}}(\epsilon) \le C_{\mathrm{diam}}(\epsilon) \le C_{\mathrm{rad}}(\epsilon/2).$$

The radius is target-centered; the diameter gives a target-uniform stopping rule. Appendix A.1 proves the proposition and the hitting-set and decision-tree characterizations.

Can the same family demand information non-monotonically? Index predictors by hidden d-bit secrets. At block horizon k, distinct secrets have distance

$$D_{d,k} = 2\big[(1-2^{-d})^k - (1-2^{1-d})^k\big].$$

The distance first grows as one trigger becomes likely, then shrinks as both triggers saturate.

Theorem 3.2 (Protocol-level prefix-trigger separation)**.** *For every $d \ge 2$ there is a class $\mathcal{H}_d$ of $3d$-state predictors with a common endpoint channel. For any nonempty finite set of aligned block horizons $\mathcal{K}$, distinct targets have protocol distance*

$$D_{d,\mathcal{K}} = \max_{k\in\mathcal{K}} D_{d,k}.$$

For every $\epsilon \ge 0$ and $0 < \delta \le 1/2$,

$$\Big(\max_s \mathrm{LGID}^{\mathcal{K}}_\epsilon(F_s),\ \mathrm{AQ}^{\mathcal{K}}_\epsilon(\mathcal{H}_d),\ \mathrm{EAC}^{\mathrm{pass,TI},\mathcal{K}}_{\epsilon,\delta}(\mathcal{H}_d)\Big) = \begin{cases} (0,0,0), & D_{d,\mathcal{K}} \le \epsilon, \\ \big(1, d, \Theta(2^d \log(1/\delta))\big), & D_{d,\mathcal{K}} > \epsilon. \end{cases}$$

A teacher presents the secret, an active learner discovers one bit per prefix query, and a target-independent passive sampler spreads mass across 2^d candidates.

Corollary 3.3 (Point versus cumulative horizons)**.** *For point evaluation $\mathcal{K} = \{k\}$, information demand can be trivial, then nontrivial, then trivial as k grows. For the nested protocol $\mathcal{K}_{\le K} = \{1,\dots,K\}$,*

$$D^{\le}_{d,K} = \max_{1\le k\le K} D_{d,k}$$

is nondecreasing, so the nontrivial requirement never disappears after the distinguishable window has been encountered.

At $k = 2^{d-1}$, $D_{d,k} > e^{-1}$. The exact cumulative maximum, construction, and matching passive bounds are in Appendix A.4.

3.2 RECOVERY: CALIBRATING EVIDENCE TO A COMPUTATION

Why can sufficient evidence still fail? Recovery asks whether the declared learner class can represent an evidence-supported computation, whether the constraints determine it, how well conditioned fitting is, and whether each local constraint receives enough exposure. If the learner class cannot express the required computation, this is a Recovery-side realizability limit; if the declared abstraction cannot express the target behavior at all, the abstraction is misspecified rather than a fifth stage.

The controlled linear specialization separates rank from conditioning. For $Y = W^\star\Phi$, $G = \Phi\Phi^\top$, and $\Phi \in \mathbb{R}^{r\times n}$:

Proposition 3.4 (Linear Recovery)**.** *The consistent solution set has dimension $d_{\text{out}}(r - \operatorname{rank}\Phi)$. If $G \succ 0$, gradient descent with step $1/\lambda_{\max}(G)$ contracts error by $1 - 1/\kappa(G)$ per step.*

For a table with M local constraints, normalized exposure $E = UB/M$ removes the confound between table size and optimizer visits:

Proposition 3.5 (Clean exposure specialization)**.** *Under independent squared-loss rows and a balanced schedule, fixed tolerance requires $U = \Theta((M/B)\log(1/\tau))$ updates.*

Exposure is not sufficient: rank, conditioning, scheduling, and parameter sharing remain distinct.

Shared parameters predict a structured early shortcut. In $f_i = b + u_i$, every example updates b, while u_i changes only when row i appears.

Proposition 3.6 (Shared-bias timescale separation)**.** *Under uniform row sampling and squared loss,*

$$\bar{m}_t = (1 - \gamma(1 + M^{-1}))^t \bar{m}_0, qquad c_t = (1 - \gamma/M)^t c_0.$$

The shared mode moves M times faster at leading order, allowing a majority sink rule to precede row-specific corrections. Appendix B.1–B.3 gives the controlled derivations; these are calibration diagnostics, not claims that one specialization fully models a Transformer.

3.3 COMPOSITION: JOINT EXECUTION UNDER DEPLOYMENT

Why is local error count insufficient? An HMC table has 64 transitions. Redirecting exactly one entry to any wrong destination gives 960 tables with $A_{\text{transition}} = 63/64$, yet survival ranges from 0.5933 to 0.8743 at horizon 16 and from 0.0110 to 0.0706 at horizon 256. The error's location determines when surviving paths encounter it.

For computations on a common declared local state space, let K_t retain only probability mass that takes another correct transition with a correct local output:

$$K_t(q, q') = \sum_a \nu_t(a \mid q)\mathbf{1}\{\delta(q,a) = q' = \widehat{\delta}(q,a),\ \widehat{r}(q,a) = r(q,a)\}.$$

Theorem 3.7 (Finite-horizon never-error survival)**.** *The probability of remaining state/output-correct for L steps is*

$$S_L = e_{q_0}^\top K_0 K_1 \cdots K_{L-1}\mathbf{1}.$$

For a stationary process, $S_L = e_{q_0}^\top K^L \mathbf{1}$.

Corollary 3.8 (Exact-Recovery closure)**.** *If transitions and outputs agree on deployment support, then $S_L = 1$ for every finite L.*

Composition is not an additional failure after exact Recovery; it describes the deployment effect of discrepancies that remain.

Never-error survival is stronger than endpoint correctness. For two deterministic machines under the same declared input process, let $z_t = (q_t, \widehat{q}_t)$, let $P_t^{(\ell)}$ be the pair-state transition at horizon ℓ, and let $m(q, \widehat{q}) = \mathbf{1}\{\rho(q) \neq \widehat{\rho}(\widehat{q})\}$.

Proposition 3.9 (Pair-state endpoint bridge)**.** *For every declared horizon ℓ,*

$$d_\ell(h, g) = e_{(q_0, \widehat{q}_0)}^\top P_0^{(\ell)} \cdots P_{\ell-1}^{(\ell)} m.$$

If the no-local-error event measured by $S_\ell(h, g)$ implies endpoint agreement, then

$$d_\ell(h, g) \leq 1 - S_\ell(h, g).$$

The pair process retains divergent trajectories that can later resynchronize; K_t is its killed synchronized restriction. Non-Markov input laws are covered by augmenting the pair state with the required history.

This bridge closes the evidence-to-deployment chain.

Theorem 3.10 (End-to-end deployment certificate)**.** *Let S be acquired evidence, choose any measurable reference $h_S \in \mathcal{V}_O(S)$, and let $\widehat{h}$ be the trained computation. Then*

$$d_{\mathcal{T}}(\widehat{h}, f^\star) \le r_{O,\mathcal{T}}(S) + d_{\mathcal{T}}(\widehat{h}, h_S).$$

When no local error implies endpoint agreement,

$$d_{\mathcal{T}}(\widehat{h}, f^\star) \le r_{O,\mathcal{T}}(S) + \sup_{\ell \in \mathcal{T}} [1 - S_\ell(\widehat{h}, h_S)].$$

Identification and Acquisition control the first term; Recovery determines the learned/reference discrepancy; joint execution evaluates its deployment consequence.

Corollary 3.11 (Probabilistic cascade)**.** *If $\Pr[r_{O,\mathcal{T}}(S_m) \le \epsilon_{\mathrm{id}}] \ge 1 - \delta_{\mathrm{acq}}$ and, conditional on this event, $\Pr[d_{\mathcal{T}}(\widehat{h}, h_{S_m}) \le \epsilon_{\mathrm{exec}}] \ge 1 - \delta_{\mathrm{rec}}$, then*

$$\Pr[d_{\mathcal{T}}(\widehat{h}, f^\star) \le \epsilon_{\mathrm{id}} + \epsilon_{\mathrm{exec}}] \ge 1 - \delta_{\mathrm{acq}} - \delta_{\mathrm{rec}}.$$

No independence assumption is required.

Finally, deployment horizon determines which symbolic repairs matter. For stationary K,

Proposition 3.12 (Path sensitivity and repair complementarity)**.**

$$\frac{\partial S_L}{\partial K_{ij}} = \sum_{t=0}^{L-1} (e_{q_0}^\top K^t)_i \, (K^{L-1-t}\mathbf{1})_j.$$

If $\Delta_1, \Delta_2 \ge 0$ entrywise and all repaired kernels are feasible row-substochastic operators, then

$$S_L(K + \Delta_1 + \Delta_2) - S_L(K + \Delta_1) - S_L(K + \Delta_2) + S_L(K) \ge 0.$$

Importance is surviving mass reaching the source state times remaining survival after the destination. Complementarity means two symbolic repairs can unlock a path jointly even when either alone has no effect. Neural parameter updates need not be entrywise local; shared parameters can alter other rows. Appendix C.1–C.2 proves these claims and retains the operator-estimation and scalar-approximation bounds.

4 Experiments: Diagnosing the Four Bottlenecks

The experiments ask whether stage-specific objects distinguish failures that endpoint accuracy merges. Full protocols, per-seed results, and scope conditions are in Appendix E.

The measurements follow the cascade. When the diagnostic class is finite, we first enumerate or exactly search its consistent alternatives to compute r_O and r_S, before considering the neural prediction. We then measure Recovery on the declared local table and evaluate Composition from the reconstructed table through K. This order prevents an optimization failure from being mislabeled as missing information, or a harmful error location from being summarized only by average local accuracy.

4.1 Acquisition: access to decisive evidence

Does the acquisition process obtain behaviorally sufficient evidence, and can a learner use it? Under the same endpoint label type, passive RPNI solves $10/10$ Random-DFA cases at $N = 8$ but none from $N = 64$, whereas membership queries solve $10/10$ at $N = 64$. On 16-state HMC, queries reach the 0.99 criterion in 297.7 ± 69.1 labels on $3/3$ seeds; passive sampling solves $0/3$ within 4096. These are learner/process comparisons, not estimates of Theorem 3.2's complexity law.

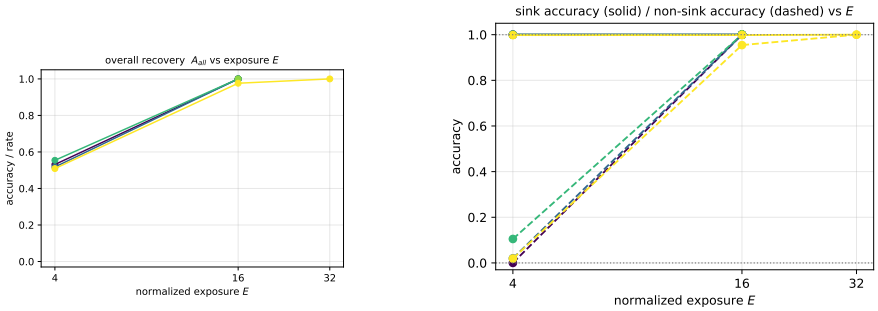


Figure 2: Bounded-Dyck Recovery. **Left:** overall table accuracy rises with normalized exposure across state counts. **Right:** low exposure preserves sink predictions while missing most non-sink transitions. The failure is a majority shortcut, not uniform noise.

Passive failure may come from ambiguous evidence or from the learner. Exact search within the declared canonical diagnostic class finds 47 evidence-ambiguity cases, five learner-only failures, and two learner successes despite ambiguity. Another 48 propagation-capped cells are undetermined and excluded from proportions. The SAT conclusion is class-relative; Appendix E.6 gives the encoding and soundness gates.

The family-aware prefix-trigger experiment asks whether behavioral evidence sufficiency has observable consequences for a neural learner; it does not estimate a neural $1 : d : 2^d$ sample-complexity law. The oracle support directly identifies the target secret. The non-oracle sufficient support omits that direct atom but still leaves a singleton version space within the known family. The two insufficient supports leave exactly two secrets or a larger ambiguous set. Because the family contains only 2^6 secrets, we enumerate it and compute

$$r_S(k) = \max_{F_t \in \mathcal{V}(S)} d_{\text{long}}(F_s, F_t).$$

Hence the oracle and sufficient supports have $r_S(k) = 0$, while either insufficient support retains radius $D_{6,k}$. This establishes what the evidence determines before testing what the Transformer learns.

At $d = 6$, oracle and sufficient supports recover the secret in $24/24$ seed–secret cases, while both insufficient supports recover it in $0/24$. In the intermediate horizon window, natural accuracy is $1.0000/0.9883/0.8075/0.6666$ for oracle, sufficient, two-hypothesis, and more ambiguous support. At $k = 160$, the two-hypothesis arm reaches 0.9270, close to the always-one baseline 0.9195, while secret Recovery remains $0/24$. Accuracy rises because the deployment criterion becomes less discriminating, not because the hidden computation has been recovered.

Sufficient support removes the information obstruction but does not guarantee robust neural execution. All 24 cases recover the secret and exhaustive one-block rule, yet one support context exhibits a localized long-horizon failure. This anomaly does not identify a particular attention or positional mechanism. Full support-fit and per-context results are in Appendix E.3.

4.2 Recovery: fitting the supplied computation

Once the complete transition table is supplied, does the learner recover it? In bounded Dyck, matched exposure controls a major scaling confound. At $E = 16$, full-table fit is exact through 512 states, 0.99992 at 1024, and 0.97636 ± 0.00924 at 2048. Increasing only the 2048-state condition to $E = 32$ restores 1.000 on $3/3$ seeds.

Here full-table Recovery means accuracy on the same declared state–symbol table used for supervision, not generalization to unseen transitions. A system with N states and alphabet Σ has $M = N|\Sigma|$ supervised rows. Normalized exposure $E = UB/M$ therefore measures the number of optimization examples per row when updates U and batch size B vary. Comparing at fixed E separates growing table size from simply giving each constraint fewer visits.

Low exposure produces a structured shortcut rather than uniform noise. Across $E = 4$ checkpoints, overall table accuracy is $A_{\text{all}} \approx 0.528$, but the predicted sink rate is $R_{\text{pred sink}} \approx 0.974$ and non-sink accuracy is only $A_{\text{nonsink}} \approx 0.036$. At $N = 32$, every entry is predicted as the sink. Higher exposure recovers the row-specific non-sink corrections, matching the shared-bias timescale separation.

4.3 Composition: state-dependent survival

Given residual local errors, can average accuracy predict long-horizon reliability? On 560 partially recovered HMC checkpoints, we compare the exact operator K, a target-visitation scalar, and the table-mean scalar $A^L_{\text{transition}}$. Their short-horizon mean absolute errors are 0.000625, 0.010, and 0.124. State dependence remains important even after the scalar receives the correct target-state visitation frequencies. These checkpoints form a structural population, not 560 independent replicates.

For each checkpoint, the transition head is evaluated on all 16×4 state–symbol pairs to reconstruct its symbolic table. Under the known uniform symbol law, this table defines K exactly. We compare $e_{q_0}^\top K^L \mathbf{1}$ with empirical probability that a rollout has never diverged from the target, for every $L \leq 256$. The visitation scalar knows how often the target visits each state but does not update that occupancy after conditioning on earlier survival; the table-mean scalar discards state visitation entirely.

This difference is not explained only by unequal Recovery levels. At exactly $A_{\text{transition}} = 0.296875$, 53 checkpoints representing 20 distinct learned tables include both zero and nonzero finite-horizon survival. One matched pair has start-state-averaged $S_{16} = 0$ versus 5.77×10^{-5}. Equal average local Recovery therefore does not determine Composition reliability; full bucket statistics are in Appendix E.11.

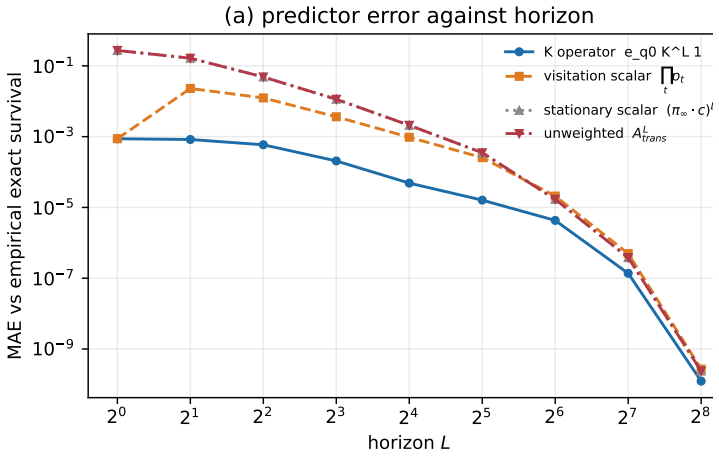


Figure 3: Prediction error across deployment horizons for 560 partially recovered HMC checkpoints. The state-dependent operator retains trajectory structure lost by both scalar baselines.

Addition is the nearly homogeneous special case: state-conditioned emitter accuracy varies by only 0.0139, and the p^L reference is within 0.021 of observed sequence accuracy. Pair-state analysis also separates never-error survival from resynchronization: HMC has $S_{256} \approx 0$ but current synchronization $C_{256} \approx 1/16$ because diverged trajectories can later meet. Details are in Appendix E.9.

4.4 Four-stage neural fingerprint

Can the diagnostics distinguish failures with similar endpoint scores? We create four stage-targeted HMC perturbations while holding the task, Transformer, optimizer, rollout path, and horizon-256 metric fixed. The perturbations change the evidence channel, acquired subset, optimization exposure, or placement of residual errors; they are diagnostic targets, not independent parameter modules.

The radii are constructed before inspecting neural performance. For each condition, r_O is the maximum horizon-256 disagreement among alternatives consistent with every atom permitted by the channel, while r_S uses only the atoms actually presented during training. We then evaluate the trained model on the full transition table and compute S^K_{256} from its reconstructed table. Identification restricts the channel so a bad alternative remains indistinguishable. Acquisition keeps a sufficient channel but withholds the decisive evidence. Recovery supplies sufficient evidence but reduces effective fitting. Composition retains high table accuracy but places the remaining error on an important trajectory.

This construction makes the signatures comparable. Perfect fit to presented examples cannot resolve a nonzero radius, because the radius concerns an alternative computation rather than the trained model. Conversely, $r_O = r_S = 0$ certifies information sufficiency only within the declared class; it does not imply that optimization recovered the target or that the residual errors compose safely.

The fingerprints separate the stages. Identification retains radius $r_O \approx 0.868$ despite perfect seen fit. Acquisition has $r_O = 0$ but acquired-evidence radius $r_S \approx 0.875$, again with perfect seen fit. Recovery has $r_O = r_S = 0$ but full-table accuracy $A_{\text{transition}} \approx 0.848$. Composition reaches $A_{\text{transition}} = 0.984375$, yet a high-impact residual error leaves operator survival $S^K_{256} \approx 0.015$. The four endpoint accuracies span only 0.0128, and each signature is stable across 15 seeds.

Stage-matched interventions produce different repairs. Restoring or acquiring decisive evidence solves the information-limited conditions; 2400 extra updates on unchanged incomplete evidence repair $0/9$ runs. With complete evidence, increased exposure solves $9/9$ Recovery runs. Ordinary full-objective training repairs the high-impact Composition error, whereas a surgical update to one local pair reduces overall table accuracy because shared parameters couple that constraint to the rest of the table. Full definitions and outcomes are in Appendix E.12.

5 Discussion

The four contracts are criterion-relative. For thresholds $(\epsilon, \tau_{\text{local}}, L, \tau_{\text{comp}})$, write A_{local} for the declared local-Recovery score. Table 1 identifies the earliest demonstrated bottleneck. The stages can coexist, and passing a local threshold is not exact Recovery; it only licenses the next diagnostic question.

First bottleneck	Diagnostic signature	Effective intervention
Identification	$r_O > \epsilon$: the allowed channel retains a bad alternative	Enrich observations or restrict the hypothesis class
Acquisition	$r_O \le \epsilon$ but $r_S > \epsilon$: decisive evidence exists but was not obtained	Reweight sampling, query adaptively, or use a curriculum
Recovery	$r_S \le \epsilon$, but $A_{\text{local}} < \tau_{\text{local}}$	Improve rank, conditioning, exposure, or parameterization
Composition	$r_S \le \epsilon$, $A_{\text{local}} \ge \tau_{\text{local}}$, but $S_L < \tau_{\text{comp}}$	Improve high-impact residual errors or rollout reliability

Table 1: The earliest unmet contract determines which resource must be addressed first.

The rule is sequential, not a claim that the stages are causally independent. Declare the protocol and tolerance ϵ, evaluate the complete channel through r_O, and only then judge acquired evidence through r_S. A small r_S licenses a Recovery audit; adequate local Recovery licenses the Composition audit. These thresholds come from deployment requirements, not from whichever model happens to pass.

Repairing the earliest failure opens the next question but does not guarantee later contracts. A decisive witness can expose an optimization problem, and improved table fit can leave a high-impact residual error. Because shared parameters also couple local updates across rows, downstream quantities should be recomputed after each intervention.

Point-horizon non-monotonicity is not a claim that every longer deployment is easier after saturation. It arises because a single late endpoint can forget an earlier distinguishable regime. Nested all-up-to-horizon protocols cannot forget it. In the neural study, sufficient support removes version-space

ambiguity but one context still fails during rollout, whereas insufficient support can score well after the point criterion saturates.

Endpoint correctness and never-error survival also answer different deployment questions. The pair-state process retains divergence and resynchronization and exactly evaluates endpoint disagreement; K_t kills a path at its first declared local error. Exact Recovery closes both criteria, but under incomplete Recovery the choice must match the deployment objective.

Several scope conditions remain. SAT conclusions are relative to the encoded class, the prefix-trigger neural study does not establish neural $1 : d : 2^d$ sample complexity, and the 560 HMC checkpoints are not independent replicates. Symbolic repair complementarity assumes feasible entrywise changes to K; shared neural parameters need not realize them locally. For black-box tasks, every stage also depends on the declared computational abstraction.

6 Conclusion

Deployment determines what finite evidence must distinguish. The prefix-trigger family makes this dependence explicit: point-horizon demand can rise and fall, while nested all-up-to-horizon demand is monotone. The version-space radius measures uncertainty left by evidence; joint execution measures the discrepancy of the learned computation.

The deployment certificate combines these terms without assuming independent failures. Family-aware Transformers separate sufficient support from saturated endpoint success, while HMC models separate average local fit from trajectory reliability and expose distinct bottlenecks behind similar scores. Length generalization depends on whether evidence identifies a deployment-safe computation, whether learning recovers one, and how its remaining discrepancies behave along the declared protocol.

AI USE STATEMENT

Generative AI tools were used for language editing and manuscript organization. The authors reviewed and revised all AI-assisted text and independently verified the mathematical statements, proofs, citations, figures, and experimental claims. The authors take responsibility for the complete content of this work, including all text and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Complete proofs and theoretical constructions are provided in the appendix. The appendix also documents the experimental protocols, diagnostic classes, evaluation criteria, seed counts, hyperparameters, intervention definitions, and scope conditions underlying the reported results. At the time of submission, we will release an anonymized repository containing the code and experimental artifacts supporting this work, including configuration files, evaluation and figure-generation scripts, and the materials needed to reproduce the reported analyses.

REFERENCES

Dana Angluin. Learning regular sets from queries and counterexamples. *Information and Computation*, 75(2):87–106, 1987.

Thomas Chen, Tengyu Ma, and Zhiyuan Li. Non-asymptotic length generalization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 8621–8682. PMLR, 2025. URL https://proceedings.mlr.press/v267/chen25ar.html.

Ying Fan, Yilun Du, Kannan Ramchandran, and Kangwook Lee. Looped transformers for length generalization. In *NeurIPS 2024 Workshop on Mathematical Reasoning and AI*, 2024. URL https://openreview.net/forum?id=PEdOdntGJG.

Sally A. Goldman and Michael J. Kearns. On the complexity of teaching. *Journal of Computer and System Sciences*, 50(1):20–31, 1995.

David Haussler and Emo Welzl. Epsilon-nets and simplex range queries. *Discrete & Computational Geometry*, 2:127–151, 1987.

Kaiying Hou, David Brandfonbrener, Sham M. Kakade, Samy Jelassi, and Eran Malach. Universal length generalization with turing programs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 23873–23893. PMLR, 2025. URL https://proceedings.mlr.press/v267/hou25a.html.

Xinting Huang, Andy Yang, Satwik Bhattamishra, Yash Sarrof, Andreas Krebs, Hattie Zhou, Preetum Nakkiran, and Michael Hahn. A formal framework for understanding length generalization in transformers. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=U49N5V51rU.

Malte Isberner, Falk Howar, and Bernhard Steffen. The open-source learnlib: A framework for active automata learning. In *Computer Aided Verification*, pp. 487–495. Springer, 2015.

Zachary Izzo, Eshaan Nichani, and Jason D. Lee. Quantitative bounds for length generalization in transformers. *arXiv preprint arXiv:2510.27015*, 2025. URL https://arxiv.org/abs/2510.27015.

Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. *arXiv preprint arXiv:2305.19466*, 2023. URL https://arxiv.org/abs/2305.19466.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*, 2021. URL https://arxiv.org/abs/2112.00114.

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

Gail Weiss, Yoav Goldberg, and Eran Yahav. Thinking like transformers. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11080–11090. PMLR, 2021. URL https://proceedings.mlr.press/v139/weiss21a.html.

Andy Yang, Pascal Bergsträßer, Georg Zetzsche, David Chiang, and Anthony W. Lin. Length generalization bounds for transformers. *arXiv preprint arXiv:2603.02238*, 2026. URL https://arxiv.org/abs/2603.02238.

Hattie Zhou, Arwen Bradley, Etai Littwin, Noam Razin, Omid Saremi, Josh Susskind, Samy Bengio, and Preetum Nakkiran. What algorithms can transformers learn? a study in length generalization. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=AssIuHnmHX.

Yongchao Zhou, Uri Alon, Xinyun Chen, Xuezhi Wang, Rishabh Agarwal, and Denny Zhou. Transformers can achieve length generalization but not robustly. *arXiv preprint arXiv:2402.09371*, 2024b. URL https://arxiv.org/abs/2402.09371.

A Identification and Acquisition

For a deployment protocol $\mathcal{T}$, write

$$\mathcal{B}_{\epsilon,\mathcal{T}}(f^\star) = \{h : d_{\mathcal{T}}(h, f^\star) > \epsilon\}, \qquad \Delta_O(h, f^\star) = \{e \in \mathcal{E}^O_{\text{train}} : O_e(h) \neq O_e(f^\star)\}.$$

A.1 Protocol metrics and certificate criteria

Proof of Proposition 3.1. For each horizon ℓ, the pointwise inclusion

$$\mathbf{1}\{h(x) \neq g(x)\} \leq \mathbf{1}\{h(x) \neq u(x)\} + \mathbf{1}\{u(x) \neq g(x)\}$$

implies the triangle inequality for d_ℓ after expectation. Symmetry and nonnegativity are immediate, so d_ℓ is a pseudometric. A supremum of pseudometrics is a pseudometric, proving the claim for $d_{\mathcal{T}}$; quotienting by zero distance makes it a metric.

If $\mathcal{T}_1 \subseteq \mathcal{T}_2$, the supremum for $\mathcal{T}_1$ ranges over fewer horizons, hence $d_{\mathcal{T}_1}(h, g) \leq d_{\mathcal{T}_2}(h, g)$. The same inequality holds for every target-centered radius and every version-space diameter. Therefore every certificate, passive-success event, or safe active leaf for $\mathcal{T}_2$ is also valid for $\mathcal{T}_1$. Taking minima, suprema over targets, and infima over common passive distributions preserves the stated monotonicity directions.

Finally, if $f^\star \in \mathcal{V}$, then

$$r_{f^\star}(\mathcal{V}) = \sup_{h \in \mathcal{V}} d(h, f^\star) \leq \sup_{h,g \in \mathcal{V}} d(h, g) = \operatorname{diam}(\mathcal{V}).$$

For every $h, g \in \mathcal{V}$, the triangle inequality gives $d(h, g) \leq d(h, f^\star) + d(g, f^\star) \leq 2r_{f^\star}(\mathcal{V})$, proving the reverse bound. A diameter-ϵ certificate is therefore a radius-ϵ certificate, while a radius-$\epsilon/2$ certificate is a diameter-ϵ certificate. Minimizing cost gives $C_{\text{rad}}(\epsilon) \leq C_{\text{diam}}(\epsilon) \leq C_{\text{rad}}(\epsilon/2)$. □

Proposition A.1 (Certificate characterization)**.** $\mathrm{LGID}^{O,\mathcal{T}}_\epsilon(f^\star)$ *is the minimum hitting-set size for*

$$\{\Delta_O(h, f^\star) : h \in \mathcal{B}_{\epsilon,\mathcal{T}}(f^\star)\}.$$

Proof. For every h and S,

$$h \in \mathcal{V}_O(S) \quad \Longleftrightarrow \quad S \cap \Delta_O(h, f^\star) = \emptyset.$$

Thus $r_{O,\mathcal{T}}(S) \leq \epsilon$ exactly when S intersects the distinguishing set of every protocol-bad hypothesis. Minimizing $|S|$ gives the result, including $+\infty$ when no finite hitting set exists. □

A.2 PASSIVE ACQUISITION BOUNDS

Proposition A.2 (Finite-class passive acquisition)**.** *Suppose $N_\epsilon < \infty$ hypotheses have protocol distance above ϵ. Let*

$$\gamma_\mu(h) = \mu(\Delta_O(h, f^\star)), \qquad \underline{\gamma}_\epsilon = \min_{h \in \mathcal{B}_{\epsilon,\mathcal{T}}(f^\star)} \gamma_\mu(h).$$

If $\underline{\gamma}_\epsilon > 0$, then

$$m \geq \frac{\log N_\epsilon + \log(1/\delta)}{\underline{\gamma}_\epsilon}$$

is sufficient for passive acquisition. If a least-observable bad hypothesis attains $\underline{\gamma}_\epsilon < 1$, then

$$m \geq \frac{\log(1/\delta)}{-\log(1 - \underline{\gamma}_\epsilon)}$$

is necessary. If $\underline{\gamma}_\epsilon = 0$, acquisition is impossible under μ.

Proof. A fixed h survives m independent observations with probability $(1 - \gamma_\mu(h))^m$. The union bound gives failure probability at most

$$N_\epsilon(1 - \underline{\gamma}_\epsilon)^m \leq N_\epsilon e^{-m\underline{\gamma}_\epsilon},$$

which proves the sufficient condition. Successful acquisition must eliminate a least-observable bad hypothesis, so its survival probability $(1 - \underline{\gamma}_\epsilon)^m$ must be at most δ, proving necessity. A zero-mass distinguishing set is missed almost surely. □

Proposition A.3 (Infinite-class passive acquisition)**.** *Let*

$$\mathcal{D}_\epsilon = \{\Delta_O(h, f^\star) : d_\mathcal{T}(h, f^\star) > \epsilon\}.$$

Assume standard measurability conditions, $\mathrm{VCdim}(\mathcal{D}_\epsilon) = v < \infty$*, and $\mu(D) \geq \gamma > 0$ for every $D \in \mathcal{D}_\epsilon$. Then*

$$\mathrm{EAC}^{\mathrm{pass}}_{\epsilon,\delta} = O\left(\frac{v \log(1/\gamma) + \log(1/\delta)}{\gamma}\right).$$

Proof. The classical VC ϵ-net theorem implies that a sample of the displayed size intersects every range of μ-mass at least γ with probability at least $1 - \delta$. It therefore hits every member of $\mathcal{D}_\epsilon$. Proposition A.1 makes that sample an identifying certificate. □

A.3 ACTIVE DECISION TREES

Proposition A.4 (Decision-tree characterization)**.** $\mathrm{AQ}^{O,\mathcal{T}}_\epsilon$ *is the minimum depth of an admissible observation tree whose reachable leaves have protocol diameter at most ϵ. If each query has at most R answers and $\mathcal{P}$ is pairwise separated by more than ϵ, then*

$$\mathrm{AQ}^{O,\mathcal{T}}_\epsilon \geq \lceil \log_R |\mathcal{P}| \rceil.$$

Proof. Unroll a deterministic adaptive policy. Each internal node is its next admissible query after the transcript leading to that node, and each outgoing edge is a possible answer. The hypotheses consistent with a root-to-node transcript are exactly the node's version space. The policy stops safely for every possible target exactly when every reachable stopping leaf has long-horizon diameter at most ϵ. Policy worst-case query count is tree depth, proving the characterization in both directions.

Now let $\mathcal{P}$ be an ϵ-packing. A safe leaf contains at most one member of $\mathcal{P}$, because two would have distance greater than its allowed diameter. A depth-T tree with at most R answers per query has at most R^T leaves. Hence $R^T \geq |\mathcal{P}|$, which gives the lower bound. □

A.4 THE PREFIX-TRIGGER SEPARATION

Construction. Fix $d \geq 2$, let $M = 2^d$, and index hypotheses by secrets $s = s_1 \cdots s_d \in \{0,1\}^d$. The common admissible short-query set is

$$\mathcal{X}_d = \bigcup_{\ell=1}^{d} \{0,1\}^\ell.$$

For $x \in \mathcal{X}_d$, define $F_s(x) = 1$ if and only if x is a prefix of s. At every aligned deployment length kd, partition the string into k independent uniform d-bit blocks and set the endpoint output to one if and only if a completed block equaled s. Thus the same predictor family supports any finite protocol of aligned block horizons.

State complexity. Use initial-prefix states $I_0, \ldots, I_{d-1}$, later-block matching states $P_1, \ldots, P_{d-1}$, mismatch states $X_1, \ldots, X_d$, and one absorbing hit state H. State I_i means that the first i input bits match the secret prefix. State P_i means that at least one nonmatching block has completed and the current block has matched its first i secret bits. State X_i means that i symbols of the current block have been read, the block already mismatched, and no earlier block matched.

From I_i, a correct next bit advances to I_{i+1}, except that the correct final bit enters H; an incorrect bit enters X_{i+1}. The P_i transitions are analogous. From X_i with $i < d$, either symbol advances to X_{i+1}. From the completed-block boundary X_d, the next bit enters P_1 or X_1 according to whether it equals s_1. State H is absorbing. Output one in every I_i and in H, and zero in every P_i and X_i.

Starting from I_0, a string of length at most d has output one exactly when it is a prefix of s. After length d, output one exactly when a completed block equaled s, including at partial later blocks. The construction has $d + (d-1) + d + 1 = 3d$ states.

Long-horizon separation. For distinct secrets s, t, disagreement occurs when at least one block equals one secret and no block equals the other. Therefore the exact pairwise distance is

$$D_{d,k} := d_k(F_s, F_t) = 2\left[\left(1 - \frac{1}{M}\right)^k - \left(1 - \frac{2}{M}\right)^k\right].$$

The event that exactly one block equals s and none equals t, together with its symmetric counterpart, gives the lower bound

$$d_k(F_s, F_t) \geq 2\frac{k}{M}\left(1 - \frac{2}{M}\right)^{k-1}.$$

The horizon dependence is exactly unimodal. Direct subtraction gives

$$D_{d,k+1} - D_{d,k} = \frac{2}{M}\left[2\left(1 - \frac{2}{M}\right)^k - \left(1 - \frac{1}{M}\right)^k\right].$$

Thus the sequence increases while

$$k < T_d := \frac{\log 2}{\log((M-1)/(M-2))}$$

and decreases after this threshold. Its integer maximum occurs at $\lceil T_d \rceil$; if T_d is an integer, both $k = T_d$ and $k = T_d + 1$ maximize the sequence. Consequently $\{k : D_{d,k} > \epsilon\}$ is a contiguous integer interval or empty.

For any nonempty finite block-horizon protocol $\mathcal{K}$, every distinct pair therefore has exact protocol distance

$$d_{\mathcal{K}}(F_s, F_t) = D_{d,\mathcal{K}} := \max_{k \in \mathcal{K}} D_{d,k}.$$

For $\mathcal{K}_{\leq K} = \{1, \ldots, K\}$, unimodality gives

$$D^{\leq}_{d,K} = D_{d,\min\{K, \lceil T_d \rceil\}}.$$

If T_d is an integer, $D_{d,T_d} = D_{d,T_d+1}$, so the displayed choice is still valid. Hence the cumulative distance is nondecreasing and becomes constant after reaching the point-horizon maximum.

The asymptotic regimes follow from the exact formula:

$$\begin{aligned} k/M \to 0 : &\quad D_{d,k} \sim 2k/M, \\ k/M \to c \in (0,\infty) : &\quad D_{d,k} \to 2(e^{-c} - e^{-2c}), \\ k/M \to \infty : &\quad D_{d,k} \to 0. \end{aligned}$$

Moreover, $T_d/M \to \log 2$ and the maximum distance tends to $1/2$. Short horizons rarely expose either trigger, intermediate horizons expose one but not the other with constant probability, and extremely long horizons make both predictors output one almost surely.

Trivial regime. If $D_{d,\mathcal{K}} \le \epsilon$, every target lies within ϵ of every other target throughout the protocol. Empty evidence therefore gives $\mathrm{LGID}_\epsilon^{\mathcal{K}} = 0$; the root version space has protocol diameter at most ϵ, so $\mathrm{AQ}_\epsilon^{\mathcal{K}} = 0$; and zero passive observations suffice.

Teacher. Suppose $D_{d,\mathcal{K}} > \epsilon$. The endpoint atom $(s, 1)$ eliminates every F_t, $t \neq s$, because no other length-d secret regards s as its prefix. Hence $\mathrm{LGID}_\epsilon^{\mathcal{K}}(F_s) \le 1$. Empty evidence retains every other F_t, all at protocol distance greater than ϵ, so $\mathrm{LGID}_\epsilon^{\mathcal{K}}(F_s) \ge 1$.

Active learner. Suppose the learner has established the prefix $p = s_1 \cdots s_{i-1}$. Query $p0$. A positive response implies $s_i = 0$; a negative response implies $s_i = 1$, because p is already known to be correct. Beginning with the empty prefix and repeating for $i = 1, \ldots, d$ identifies s in d binary queries. Conversely, the 2^d targets are pairwise more than ϵ apart, and every endpoint query has two answers. Proposition A.4 gives $\mathrm{AQ}_\epsilon^{\mathcal{K}} \ge \log_2(2^d) = d$. Therefore $\mathrm{AQ}_\epsilon^{\mathcal{K}} = d$.

Target-independent minimax passive sampler. Let μ be any distribution on $\mathcal{X}_d$ fixed independently of the target. Partition the length-d strings into 2^{d-1} sibling pairs

$$P_u = \{u0, u1\}, \qquad u \in \{0,1\}^{d-1}.$$

Because these pairs are disjoint,

$$\sum_u \mu(P_u) = \mu(\{0,1\}^d) \le 1.$$

Hence some $P_u = \{s, s'\}$ has $\mu(P_u) \le 2^{1-d}$. Choose F_s as the target. The siblings have identical prefix answers at every length below d; at length d, they differ only on the two strings s and s'. Thus any transcript that misses P_u leaves $F_{s'}$ in the version space. The long-horizon calculation above gives $d_{\mathcal{K}}(F_s, F_{s'}) = D_{d,\mathcal{K}} > \epsilon$, so missing the pair is a failure event. After m independent observations its probability is

$$(1 - \mu(P_u))^m \ge (1 - 2^{1-d})^m.$$

Success probability at least $1 - \delta$ for this target therefore requires

$$m \ge \left\lceil \frac{\log(1/\delta)}{-\log(1 - 2^{1-d})} \right\rceil.$$

Because the worst-case target may depend on the common distribution μ, this inequality gives a lower bound on $\inf_\mu \sup_s \mathrm{EAC}_{\epsilon,\delta}^{\mathrm{pass},\mathcal{K}}(F_s; \mu)$.

For the matching upper bound, let $\mu^\star$ be uniform on the length-d strings and assign zero mass to shorter strings. The target string s has mass 2^{-d}, and observing $(s, 1)$ eliminates every other member of $\mathcal{H}_d$. Hence, simultaneously for every target,

$$\mathrm{EAC}_{\epsilon,\delta}^{\mathrm{pass},\mathcal{K}}(f_s) \le \left\lceil \frac{\log(1/\delta)}{-\log(1 - 2^{-d})} \right\rceil.$$

The bounds are respectively $\Omega(2^d \log(1/\delta))$ and $O(2^d \log(1/\delta))$. Together with the trivial regime, this proves Theorem 3.2. □

Proof of Corollary 3.3. For point evaluation, the unimodality calculation shows that $D_{d,k}$ first rises and then converges to zero. Theorem 3.2 therefore gives the possible trivial–nontrivial–trivial pattern. For cumulative evaluation, $D^{\leq}_{d,K}$ is nondecreasing, and applying the theorem to each nested $\mathcal{K}_{\leq K}$ shows that the nontrivial regime persists once entered.

For the stated intermediate witness, set $k = M/2$ and $n = M/2 \geq 2$. The lower bound above becomes

$$D_{d,k} \geq 2\frac{k}{M}\left(1 - \frac{2}{M}\right)^{k-1} = \left(1 - \frac{1}{n}\right)^{n-1} > e^{-1}.$$

The last inequality follows from $\log(1-x) > -x/(1-x)$ with $x = 1/n$. Thus $D_{d,2^{d-1}} > e^{-1}$. □

Uniform-on-all-short-strings specialization. In the nontrivial protocol regime $D_{d,\mathcal{K}} > \epsilon$, let μ_d instead be uniform on $\mathcal{X}_d$, whose size is $U_d = \sum_{\ell=1}^{d} 2^\ell = 2^{d+1} - 2$. Hitting s identifies the target, while missing $\{s, s'\}$ leaves its final-bit sibling consistent. Therefore

$$\frac{\log(1/\delta)}{-\log(1 - 2/U_d)} \leq \mathrm{EAC}^{\mathrm{pass}}_{\epsilon,\delta}(F_s; \mu_d) \leq \left\lceil \frac{\log(1/\delta)}{-\log(1 - 1/U_d)} \right\rceil.$$

Using

$$p \leq -\log(1-p) \leq \frac{p}{1-p}$$

gives the explicit comparison

$$\frac{U_d - 2}{2}\log\frac{1}{\delta} \leq \mathrm{EAC}^{\mathrm{pass}}_{\epsilon,\delta}(F_s; \mu_d) \leq U_d \log\frac{1}{\delta} + 1.$$

This explicit distribution remains useful for the per-target calculation, but the minimax theorem above is stronger because it optimizes over all common passive distributions.

A.5 Persistent-pattern specialization

For $s \in \{0,1\}^d$, let $f_s(x) = 1$ when s has appeared as a substring of x, and include the constant-zero hypothesis h_0. A deterministic pattern matcher stores the length of the longest suffix that is a prefix of s; d prefix states and one absorbing accepting state suffice.

On length-d inputs, the target is positive only at $s^\star$. The atom $(s^\star, 1)$ eliminates h_0 and every f_s with $s \neq s^\star$, so LGID is one. The empty set leaves h_0. For long length $d2^d$, partition the input into 2^d independent uniform blocks. If any block equals $s^\star$, the target and h_0 disagree, so

$$d_{\mathrm{long}}(h_0, f_{s^\star}) \geq 1 - (1 - 2^{-d})^{2^d} \geq 1 - e^{-1}.$$

Under uniform passive length-d sampling, h_0 is eliminated if and only if $s^\star$ is observed. Thus success probability is $1 - (1-2^{-d})^m$, and solving for confidence $1 - \delta$ gives

$$\mathrm{EAC}^{\mathrm{pass}}_{\epsilon,\delta} = \left\lceil \frac{\log\delta}{\log(1 - 2^{-d})} \right\rceil = \Theta\left(2^d \log\frac{1}{\delta}\right). \tag{1}$$

□

A.6 Additional structural consequences

Proposition A.5 (Finite versus infinite LGID)**.** *For a fixed protocol $\mathcal{T}$, if $\mathcal{H}$ is finite then $\mathrm{LGID}^{O,\mathcal{T}}_{\epsilon} < \infty$ if and only if every $h \in \mathcal{B}_{\epsilon,\mathcal{T}}(f^\star)$ has a nonempty distinguishing set. For an infinite class, finiteness requires the distinguishing-set family to admit a finite hitting set; pairwise distinguishability alone is insufficient.*

Proof. An empty distinguishing set cannot be hit. Conversely, for a finite class choose one atom from every nonempty bad-hypothesis distinguishing set and take their union. The infinite-class statement follows directly from Proposition A.1. □

Proposition A.6 (Monotonicity and observation refinement)**.** *Enlarging the hypothesis class cannot decrease LGID; increasing ϵ cannot increase it. If every atom of O_1 has an equal-cost simulation in O_2 such that agreement under O_2 implies agreement under O_1, then $\mathrm{LGID}^{O_2}_{\epsilon} \leq \mathrm{LGID}^{O_1}_{\epsilon}$.*

Proof. The first two claims follow because the family of bad distinguishing sets respectively grows and shrinks. For refinement, simulate every atom of an O_1 certificate in O_2. The resulting O_2 version space is a subset of the original O_1 version space at no greater cost. □

B RECOVERY

B.1 RANK AND CONDITIONING

Proof of Proposition 3.4. The evidence equation is $(W - W^\star)\Phi = 0$. If $k = \operatorname{rank}\Phi$, each row of $W - W^\star$ may be any linear functional in the $(r-k)$-dimensional null space of $\Phi^\top$. The rows are independent, giving affine dimension $d_{\text{out}}(r-k)$.

Let U_S be the span of the evidence features. If $U_S = U_{\text{train}}$, consistency implies agreement throughout the behaviorally faithful admissible span and hence zero long-horizon error. Conversely, if U_S is a proper subspace, choose $u \in U_{\text{train}} \setminus U_S$. A linear functional exists that vanishes on U_S but not on u. Adding it to one row of $W^\star$ produces a consistent parameter that differs on U_{train}, hence has positive long-horizon error. A smallest spanning subset has $\dim U_{\text{train}}$ vectors.

For the optimization statement,

$$\nabla L(W) = (W - W^\star)G, \qquad W_{t+1} - W^\star = (W_t - W^\star)(I - \gamma G).$$

With $\gamma = 1/\lambda_{\max}(G)$, positive definiteness gives

$$\|I - \gamma G\|_2 = 1 - \frac{\lambda_{\min}(G)}{\lambda_{\max}(G)} = 1 - \frac{1}{\kappa(G)}.$$

Iteration proves the displayed rate. If G is singular, a nonzero vector in its null space is annihilated by $\Phi^\top$, leaving an unidentified direction. □

For additive updates, a witness for symbol a contributes the standard-basis feature e_a, so one independent witness per symbol is necessary and sufficient. For scalar affine updates, $\phi(q, a) = e_a \otimes (q, 1)$; two source states are independent within a symbol block exactly when they are distinct. Multivariate affine updates require $d + 1$ affinely independent source states per symbol.

The distinction is strict even among identifying evidence systems. Take

$$\Phi_1 = I_2, \qquad \Phi_\eta = \begin{pmatrix} 1 & 0 \\ 0 & \eta \end{pmatrix}, \quad 0 < \eta < 1.$$

Both matrices have rank two and therefore the same exact Identification status. Yet $\kappa(\Phi_1\Phi_1^\top) = 1$, whereas $\kappa(\Phi_\eta\Phi_\eta^\top) = \eta^{-2}$, which is arbitrarily large as $\eta \downarrow 0$. Thus rank controls whether Recovery is possible from the evidence, while conditioning can make gradient-based Recovery arbitrarily slow.

B.2 CLEAN EXPOSURE-NORMALIZED TABLE RECOVERY

Proof of Proposition 3.5. Write $e_i^{(t)} = \theta_i^{(t)} - \theta_i^\star$. A visit to row i performs gradient descent on $\frac{1}{2}\|e_i\|^2$, so

$$e_i \longmapsto e_i - \eta e_i = (1 - \eta)e_i.$$

A balanced complete table pass visits every row exactly once. After E such passes,

$$e_i^{(E)} = (1 - \eta)^E e_i^{(0)} \quad \text{for every } i,$$

and taking the maximum block norm yields

$$\|\theta^{(E)} - \theta^\star\|_\infty = (1 - \eta)^E \|\theta^{(0)} - \theta^\star\|_\infty.$$

For initialization scale $R = \|\theta^{(0)} - \theta^\star\|_\infty$, tolerance $\|\theta^{(E)} - \theta^\star\|_\infty \le \tau$ is equivalent to

$$E \ge \frac{\log(R/\tau)}{-\log(1 - \eta)}.$$

With fixed η and fixed R, this is $\Theta(\log(1/\tau))$, independent of M. A minibatch schedule that processes B rows per optimizer update requires $UB = EM$ row visits, giving

$$U_\tau = \Theta\left(\frac{M}{B}\log\frac{1}{\tau}\right).$$

□

B.3 SHARED-BIAS MAJORITY SHORTCUT

Proof of Proposition 3.6. Write the score and error vectors as $h_t = b_t\mathbf{1} + u_t$ and $e_t = h_t - y$. If row $I_t = i$ is sampled, one SGD step gives

$$b_{t+1} = b_t - \gamma e_{i,t}, \qquad u_{j,t+1} = u_{j,t} - \gamma\mathbf{1}\{j = i\}e_{i,t}.$$

Conditional on e_t, uniform row sampling yields

$$\mathbb{E}[e_{t+1} \mid e_t] = e_t - \gamma\bar{e}_t\mathbf{1} - \frac{\gamma}{M}e_t, \qquad \bar{e}_t = \frac{1}{M}\mathbf{1}^\top e_t.$$

Linearity closes the recurrence for $m_t = \mathbb{E}[e_t]$:

$$m_{t+1} = \left(1 - \frac{\gamma}{M}\right) m_t - \gamma\bar{m}_t\mathbf{1}.$$

Taking the row mean and subtracting it from the vector give

$$\bar{m}_{t+1} = \left(1 - \gamma(1 + M^{-1})\right)\bar{m}_t, \qquad c_{t+1} = \left(1 - \frac{\gamma}{M}\right) c_t.$$

Iteration proves the first display in the proposition.

Now suppose $y_i \in \{0, 1\}$, $M^{-1}\sum_i y_i = \pi$, and $h_0 = 0$. Then $m_0 = -y$, $\bar{m}_0 = -\pi$, and $c_0 = \pi\mathbf{1} - y$. With $\alpha = 1 - \gamma(1 + M^{-1})$ and $\beta = 1 - \gamma/M$,

$$\mathbb{E}[h_t] = y + m_t = (1 - \beta^t)y + \pi(\beta^t - \alpha^t)\mathbf{1}.$$

Moreover, every coordinate satisfies

$$|\mathbb{E}[h_{i,t}] - \pi| \le (1 - \beta^t) + \pi\alpha^t.$$

Therefore, if $\pi > 1/2$ and $(1 - \beta^t) + \pi\alpha^t < \pi - 1/2$, thresholding the expected scores at $1/2$ predicts the sink label for every row. For fixed γ and $\pi > 1/2$, such an intermediate window exists for sufficiently large M: first choose t so that α^t is small, then increase M so that β^t remains near one. Since $1 - \alpha = \Theta(\gamma)$ and $1 - \beta = \gamma/M$, the shared and contrast timescales are respectively $O(1/\gamma)$ and $O(M/\gamma)$. □

The proposition concerns the expected iterate of a deliberately minimal shared-parameter model. It isolates how an every-sample global update can precede corrections that receive only one row's samples.

C COMPOSITION

C.1 COMPOSITION SURVIVAL

Proof of Theorem 3.7. Define

$$p_t(q) = \Pr[\text{the first } t \text{ transitions and outputs are correct and } q_t = q].$$

While the executions are synchronized at q, the probability at step t that the target moves to q', the learned transition follows it, and the local output is correct is exactly $K_t(q, q')$. Therefore

$$p_{t+1} = p_t K_t, \qquad p_0 = e_{q_0}^\top.$$

Induction and summation over the terminal state give

$$S_L = e_{q_0}^\top K_0 K_1 \cdots K_{L-1}\mathbf{1}.$$

If correctness additionally requires a terminal readout not already included in the stepwise indicator, replace $\mathbf{1}$ by $c(q) = \mathbf{1}\{\rho(q) = \widehat{\rho}(q)\}$. For a stationary input process, every K_t equals a common K, recovering $e_{q_0}^\top K^L\mathbf{1}$.

If K is primitive, Perron–Frobenius theory gives positive left and right eigenvectors w, v, a simple leading eigenvalue $\rho(K) > 0$, and

$$\rho(K)^{-L}K^L \longrightarrow \frac{vw^\top}{w^\top v}.$$

Multiplying by $e_{q_0}^\top$ and $\mathbf{1}$ yields

$$\frac{S_L}{\rho(K)^L} \longrightarrow \frac{(e_{q_0}^\top v)(w^\top\mathbf{1})}{w^\top v} > 0.$$

□

Proof of Corollary 3.8. Let

$$P_t(q, q') = \sum_a \nu_t(a \mid q)\mathbf{1}\{\delta(q, a) = q'\}$$

be the target transition operator. Exact transitions and stepwise outputs on every event of positive deployment probability make the defining indicator for K_t equal the one for P_t, so $K_t = P_t$. Each P_t is row-stochastic, hence $P_t\mathbf{1} = \mathbf{1}$, and

$$S_L = e_{q_0}^\top P_0 \cdots P_{L-1}\mathbf{1} = 1.$$

□

Proposition C.1 (Operator perturbation)**.** *If row-substochastic $K_t, \widehat{K}_t$ satisfy $\|\widehat{K}_t - K_t\|_\infty \le \eta_t$, then*

$$|\widehat{S}_L - S_L| \le \sum_{t=0}^{L-1} \eta_t.$$

In the stationary case, $\|\widehat{K} - K\|_\infty \le \eta$ gives $|\widehat{S}_L - S_L| \le L\eta$.

Proof. The matrix-product difference telescopes as

$$\widehat{K}_0 \cdots \widehat{K}_{L-1} - K_0 \cdots K_{L-1} = \sum_{t=0}^{L-1} \widehat{K}_0 \cdots \widehat{K}_{t-1}(\widehat{K}_t - K_t)K_{t+1} \cdots K_{L-1},$$

with empty products interpreted as the identity. A row-substochastic matrix A satisfies $\|A\|_\infty \le 1$, so submultiplicativity gives

$$\left\|\widehat{K}_0 \cdots \widehat{K}_{L-1} - K_0 \cdots K_{L-1}\right\|_\infty \le \sum_{t=0}^{L-1} \eta_t.$$

Finally, $|e_{q_0}^\top A\mathbf{1}| \le \|A\|_\infty$. The stationary claim follows by setting every $\eta_t = \eta$. □

Corollary C.2 (State-conditional finite-sample estimate)**.** *In the stationary setting, suppose each state q receives n_q independent one-step local evaluations under the deployment symbol law. Record the successful destination q' when the target and learned transition/output agree, and a failure symbol otherwise; let $\widehat{K}(q, q')$ be the empirical successful-transition frequency. With $n_{\min} = \min_q n_q$, probability at least $1 - \delta$ gives*

$$\|\widehat{K} - K\|_\infty \le |Q|\sqrt{\frac{\log(2|Q|^2/\delta)}{2n_{\min}}},$$

and consequently

$$|\widehat{S}_L - S_L| \le L|Q|\sqrt{\frac{\log(2|Q|^2/\delta)}{2n_{\min}}}.$$

Proof. For each ordered pair (q, q'), the empirical entry is an average of independent Bernoulli indicators. Hoeffding's inequality and a union bound over the $|Q|^2$ entries imply that all entrywise errors are at most $\sqrt{\log(2|Q|^2/\delta)/(2n_{\min})}$ simultaneously. Summing across any row gives the displayed operator-norm bound, and Proposition C.1 gives the survival bound. This deliberately generic result is not an optimal sample-complexity claim. □

Survival-conditioned occupancy. Let E_t be the event that execution has remained locally correct through the first t steps, and let

$$\pi_t(q) = \Pr(q_t = q), \qquad w_t(q) = \Pr(E_t \mid q_t = q), \qquad c_t = K_t\mathbf{1}.$$

The value of w_t on a zero-mass state is arbitrary. The surviving mass satisfies

$$p_t(q) = \Pr(E_t, q_t = q) = \pi_t(q)w_t(q), \qquad S_t = \sum_q p_t(q) = \mathbb{E}_{\pi_t}[w_t].$$

When $S_t > 0$, its normalized occupancy is therefore

$$\widetilde{\pi}_t(q) = \frac{p_t(q)}{S_t} = \frac{\pi_t(q) w_t(q)}{S_t}.$$

Write $b_t = \pi_t^\top c_t$ and $S_L^{\mathrm{visit}} = \prod_{t=0}^{L-1} b_t$.

Proposition C.3 (Survival–visitation covariance identity)**.** *For every t, including after extinction,*

$$S_{t+1} - S_t b_t = \mathrm{Cov}_{\pi_t}(w_t, c_t).$$

Consequently,

$$S_L - S_L^{\mathrm{visit}} = \sum_{t=0}^{L-1} \mathrm{Cov}_{\pi_t}(w_t, c_t) \prod_{j=t+1}^{L-1} b_j.$$

If $K_t \mathbf{1} = \alpha_t \mathbf{1}$, then $S_L = \prod_{t<L} \alpha_t$ exactly.

Proof. Because $c_t(q)$ is the conditional probability of surviving the next local step from synchronized state q,

$$S_{t+1} = \mathbb{E}_{\pi_t}[w_t c_t].$$

Consequently, for every t, including after extinction,

$$S_{t+1} - S_t b_t = \mathbb{E}_{\pi_t}[w_t c_t] - \mathbb{E}_{\pi_t}[w_t]\mathbb{E}_{\pi_t}[c_t] = \mathrm{Cov}_{\pi_t}(w_t, c_t).$$

If $S_t > 0$, division by S_t gives

$$\widetilde{\pi}_t^\top c_t - b_t = \frac{\mathrm{Cov}_{\pi_t}(w_t, c_t)}{S_t}.$$

The scalar recursion admits the zero-safe telescoping identity

$$S_L - \prod_{t=0}^{L-1} b_t = \sum_{t=0}^{L-1} (S_{t+1} - S_t b_t) \prod_{j=t+1}^{L-1} b_j.$$

Substitution proves the displayed covariance decomposition. Equivalently, before any extinction, set $a_t = \widetilde{\pi}_t^\top c_t = S_{t+1}/S_t$ and apply the standard product identity to $\prod_t a_t - \prod_t b_t$; the factors $\prod_{i<t} a_i = S_t$ cancel the denominator in the covariance formula. If extinction occurs, then $w_t = 0$ π_t-almost surely thereafter, so every later covariance is zero and the zero-safe identity remains valid.

Finally, if $K_t \mathbf{1} = \alpha_t \mathbf{1}$, then every row survival is α_t. Hence $S_{t+1} = \alpha_t S_t$, and induction from $S_0 = 1$ yields $S_L = \prod_{t<L} \alpha_t$, independently of the initial state. □

Corollary C.4 (Heterogeneity–occupancy bound)**.** *For the positive-survival times,*

$$|S_L - S_L^{\mathrm{visit}}| \le \frac{1}{2} \sum_{\substack{0 \le t < L \\ S_t > 0}} S_t \left(\prod_{j=t+1}^{L-1} b_j \right) \mathrm{osc}(c_t) \|\widetilde{\pi}_t - \pi_t\|_1 \le \frac{1}{2} \sum_{\substack{0 \le t < L \\ S_t > 0}} \mathrm{osc}(c_t) \|\widetilde{\pi}_t - \pi_t\|_1.$$

Proof. Let $x_t = \widetilde{\pi}_t - \pi_t$. Since $x_t^\top \mathbf{1} = 0$, subtracting the midpoint $m_t = (\max_q c_t(q) + \min_q c_t(q))/2$ gives

$$|\widetilde{\pi}_t^\top c_t - \pi_t^\top c_t| = |x_t^\top (c_t - m_t \mathbf{1})| \le \frac{1}{2} \mathrm{osc}(c_t) \|x_t\|_1.$$

Multiply by S_t, use $|\mathrm{Cov}_{\pi_t}(w_t, c_t)| = S_t |\widetilde{\pi}_t^\top c_t - b_t|$, and apply Proposition C.3. All weights in the first sum are at most one, proving the second inequality. □

Thus small state-wise survival oscillation or a small survival-conditioned occupancy shift is sufficient for scalar accuracy. Neither is necessary: the shift can be orthogonal to c_t, and signed covariance contributions can cancel across time.

C.2 PAIR-STATE RESYNCHRONIZATION

Let $z_t = (q_t, \widehat{q}_t)$. For the input process associated with horizon ℓ, conditional on the current pair state define

$$P_t^{(\ell)}((q,\widehat{q}),(q',\widehat{q}')) = \sum_a \nu_t^{(\ell)}(a \mid q,\widehat{q})\mathbf{1}\{\delta(q,a) = q',\ \widehat{\delta}(\widehat{q},a) = \widehat{q}'\}.$$

If the input law is not Markov in $(q_t, \widehat{q}_t)$, augment the pair state by the finite history needed to represent its conditional law. This does not change the endpoint event.

Proof of Proposition 3.9. The initial pair distribution is $e_{(q_0,\widehat{q}_0)}^\top$, and the Markov recursion gives

$$\Pr(z_\ell = \cdot) = e_{(q_0,\widehat{q}_0)}^\top P_0^{(\ell)} \cdots P_{\ell-1}^{(\ell)}.$$

Let

$$m(q,\widehat{q}) = \mathbf{1}\{\rho(q) \neq \widehat{\rho}(\widehat{q})\}.$$

Summing terminal pair-state mass against m is exactly the endpoint disagreement probability, proving

$$d_\ell(h,g) = e_{(q_0,\widehat{q}_0)}^\top P_0^{(\ell)} \cdots P_{\ell-1}^{(\ell)} m.$$

Let E_ℓ be the event counted by $S_\ell(h,g)$. Under the stated assumption, E_ℓ implies equal endpoint outputs, so the mismatch event is contained in E_ℓ^c. Therefore $d_\ell(h,g) \le \Pr(E_\ell^c) = 1 - S_\ell(h,g)$. □

For same-state-space machines, let $D = \{(q,q) : q \in Q\}$ be the synchronized diagonal. Restrict the pair chain to transitions from D back into D, delete mass that leaves D, and delete diagonal transitions with an incorrect local output. Identifying (q,q) with q yields K. Thus P retains divergent paths that may resynchronize, while K kills a path at its first declared local error.

For completeness, with

$$d(q,\widehat{q}) = \mathbf{1}\{q = \widehat{q}\}, \qquad g(q,\widehat{q}) = 1 - m(q,\widehat{q}),$$

the current-synchronization and endpoint-accuracy quantities are

$$C_L = e_{(q_0,\widehat{q}_0)}^\top P_0 \cdots P_{L-1} d, \qquad A_L^{\mathrm{out}} = e_{(q_0,\widehat{q}_0)}^\top P_0 \cdots P_{L-1} g.$$

Proof of Theorem 3.10. Because $d_{\mathcal{T}}$ is a pseudometric and $h_S \in \mathcal{V}_O(S)$,

$$d_{\mathcal{T}}(\widehat{h}, f^\star) \le d_{\mathcal{T}}(\widehat{h}, h_S) + d_{\mathcal{T}}(h_S, f^\star) \le d_{\mathcal{T}}(\widehat{h}, h_S) + r_{O,\mathcal{T}}(S).$$

Applying Proposition 3.9 at every declared horizon and taking the supremum gives

$$d_{\mathcal{T}}(\widehat{h}, h_S) \le \sup_{\ell \in \mathcal{T}} [1 - S_\ell(\widehat{h}, h_S)],$$

which proves the conservative survival form. □

Proof of Corollary 3.11. Let

$$A = \{r_{O,\mathcal{T}}(S_m) \le \epsilon_{\mathrm{id}}\}, \qquad B = \{d_{\mathcal{T}}(\widehat{h}, h_{S_m}) \le \epsilon_{\mathrm{exec}}\}.$$

The assumptions give $\Pr(A) \ge 1 - \delta_{\mathrm{acq}}$ and $\Pr(B \mid A) \ge 1 - \delta_{\mathrm{rec}}$. Hence

$$\Pr(A \cap B) = \Pr(A)\Pr(B \mid A) \ge (1 - \delta_{\mathrm{acq}})(1 - \delta_{\mathrm{rec}}) \ge 1 - \delta_{\mathrm{acq}} - \delta_{\mathrm{rec}}.$$

On $A \cap B$, Theorem 3.10 gives the claimed error bound. No independence is used. □

Proof of Proposition 3.12. For the matrix unit E_{ij}, differentiating the product gives

$$\frac{d}{d\eta}(K + \eta E_{ij})^L\Big|_{\eta=0} = \sum_{t=0}^{L-1} K^t E_{ij} K^{L-1-t}.$$

Multiplication by $e_{q_0}^\top$ and $\mathbf{1}$ yields

$$\frac{\partial S_L}{\partial K_{ij}} = \sum_{t=0}^{L-1} (e_{q_0}^\top K^t)_i \, (K^{L-1-t}\mathbf{1})_j,$$

where the two scalar factors are multiplied.

For complementarity, $F(X) = e_{q_0}^\top X^L \mathbf{1}$ is a polynomial in the entries of X with nonnegative coefficients, by its length-L path expansion. Define $\phi(s,t) = F(K + s\Delta_1 + t\Delta_2)$ for $s, t \in [0,1]$. Since K, Δ_1, Δ_2 are entrywise nonnegative, every mixed partial contributing to $\partial_s \partial_t \phi$ is nonnegative. Therefore

$$\begin{aligned} &F(K + \Delta_1 + \Delta_2) - F(K + \Delta_1) - F(K + \Delta_2) + F(K) \\ &\quad = \int_0^1 \int_0^1 \partial_s \partial_t \phi(s,t)\, ds\, dt \geq 0. \end{aligned}$$

Row-substochastic feasibility is needed for the repaired matrices to remain survival operators, not for the polynomial inequality itself. □

Minimal complementarity example. Let $K = 0$ on states $\{0, 1, 2\}$, take horizon $L = 2$, and define $\Delta_A(0,1) = 1$ and $\Delta_B(1,2) = 1$, with all other entries zero. Each individual and combined kernel is row-substochastic. Starting from state zero,

$$S_2(K) = S_2(K + \Delta_A) = S_2(K + \Delta_B) = 0, \qquad S_2(K + \Delta_A + \Delta_B) = 1.$$

The two repairs jointly restore a complete surviving path, although neither has isolated value.

Equal local recovery, different survival. Let $Q = \{0,1\}$, $\Sigma = \{a, b\}$, and draw symbols uniformly. For both source states, let $\delta(q, a) = 0$ and $\delta(q, b) = 1$. Learner A is wrong only on $(0, a)$, while learner B is wrong only on $(0, b)$. Both have unweighted transition-table recovery $3/4$, but

$$K_A = \begin{pmatrix} 0 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}, \qquad K_B = \begin{pmatrix} 1/2 & 0 \\ 1/2 & 1/2 \end{pmatrix}.$$

Thus

$$\rho(K_A) = \frac{1 + \sqrt{5}}{4} \approx 0.809, \qquad \rho(K_B) = \frac{1}{2}.$$

Starting from state zero, the corresponding finite-horizon survivals are

$$S_{16}^A \approx 2.44 \times 10^{-2}, \qquad S_{16}^B = 2^{-16} \approx 1.53 \times 10^{-5},$$

and

$$S_{32}^A \approx 8.21 \times 10^{-4}, \qquad S_{32}^B = 2^{-32} \approx 2.33 \times 10^{-10}.$$

Thus equal average local Recovery can have sharply different finite-horizon Composition reliability; the location of the error, not just its count, controls survival.

Independent-error special case. If every target transition survives independently with the same probability p, then $K = pP$, where P is the target Markov transition matrix. Because $P^L \mathbf{1} = \mathbf{1}$,

$$S_L = e_{q_0}^\top (pP)^L \mathbf{1} = p^L.$$

D Finite-State and Stack Specializations

Consider deterministic Moore machines $M = (Q, \Sigma, \delta, q_0, \rho)$ with endpoint behavior $f_M(x) = \rho(\delta_M^*(q_0, x))$.

Proposition D.1 (Short-endpoint obstruction). *If some $h \in \mathcal{H}$ agrees with $f^\star$ on every admissible endpoint through L_{train} but has positive long-horizon error, then $\mathrm{LGID}_0^{\text{end}} = +\infty$ for that bounded endpoint channel.*

Theorem D.2 (Bounded-state distinguishing horizon)**.** *If two inequivalent deterministic Moore machines have at most n states each, a distinguishing word has length at most $n^2 - 1$. If all endpoints through that length are admissible, then*

$$\mathrm{LGID}_0^{\mathrm{end}} \leq \sum_{j=0}^{n^2-1} |\Sigma|^j.$$

Proposition D.3 (Arbitrarily delayed alternatives)**.** *If for every finite L the class contains h_L that agrees with $f^\star$ through length L but has positive long-horizon error, then $\mathrm{LGID}_0^{\mathrm{end}} = +\infty$, even if arbitrary finite endpoint queries are admissible.*

Fix the target representation and readout. Let E_{reach} be the state–symbol edges used with positive probability under D_{long}, and let E_{sep} contain edges for which the class includes a positive-error hypothesis changing only that transition.

Theorem D.4 (Local-table transition LGID)**.** *For a fixed-representation local transition-table class whose unit-cost witness channel contains every edge in E_{reach},*

$$|E_{\mathrm{sep}}| \leq \mathrm{LGID}_0^{\mathrm{trans}} \leq |E_{\mathrm{reach}}|.$$

Equality holds when every target-reachable edge is behaviorally separating.

Proofs. Proposition D.1 follows because the stated bad hypothesis has an empty distinguishing set. For Theorem D.2, run two machines synchronously. A shortest distinguishing word cannot repeat a product state along its prefixes, since removing the intervening cycle would give a shorter word. It therefore has length at most $n^2 - 1$. Taking all words through that length hits every behaviorally inequivalent alternative. Proposition D.3 follows because every finite evidence set has a maximum query length.

For Theorem D.4, witnessing every edge in E_{reach} forces any consistent local table to follow the target on almost every evaluated trajectory, giving the upper bound. For each $e \in E_{\mathrm{sep}}$, the corresponding single-edge alternative can be eliminated only by witnessing e, giving the lower bound. □

Corollary D.5 (Bounded-stack reduction)**.** *A deterministic stack machine with control states Q, stack alphabet Γ, and maximum depth B has at most*

$$|Q| \sum_{j=0}^{B} |\Gamma|^j$$

configurations. It can be unfolded into a finite-state machine, so Theorems D.2 and D.4 apply to its bounded configuration graph.

E Experimental Protocols and Full Results

E.1 Reporting and seed semantics

The acquisition runs vary both the task instance and learner randomness. The architecture comparison fixes task seed 0 and varies initialization and training randomness. Aggregate comparisons therefore use the same task family, specification, and horizons, but not necessarily the same task instances; only seed-0 rows are exact instance matches. Unless explicitly noted, tabulated uncertainty uses sample standard deviation ($ddof = 1$).

E.2 Passive EAC validation

The persistent-pattern study evaluates $d \in \{4, 6, 8, 10, 12, 14, 16\}$. For the 95% quantile,

$$\log_2 \widehat{m}_{0.95}(d) = a + bd, \qquad b = 1.0016 \pm 0.0010,$$

with 95% confidence interval $[0.9990, 1.0043]$ and $R^2 = 0.999995$. The empirical-to-exact quantile ratio remains within about 0.8%. Normalized costs and the theoretical confidence term are:

δ	0.1	0.05	0.01
$\widehat{m}/2^d$	2.30	3.00	4.60
$\log(1/\delta)$	2.303	2.996	4.605

For $\mu_\alpha = (1-\alpha)\mu_{\text{uniform}} + \alpha\delta_{s^\star}$, the decisive mass is $p_\alpha = (1-\alpha)2^{-d} + \alpha$. Empirical acquisition costs track

$$m^\star_\alpha = \left\lceil \frac{\log \delta}{\log(1-p_\alpha)} \right\rceil$$

within roughly 2.3%, with cost reductions from $5.1\times$ to more than $6700\times$.

E.3 Family-Aware Prefix-Trigger Neural Experiment

This experiment uses the same prefix-trigger family $\mathcal{H}_d$ as Theorem 3.2, but asks a neural Recovery question rather than estimating the theorem's sample complexity. A single Standard Transformer architecture is trained in an episodic, family-aware setting; training and evaluation secrets are disjoint. Its positional representation is defined at every evaluated length. No Looped Transformer or GRU is part of this experiment. The primary condition is $d = 6$, with eight held-out evaluation secrets and three neural seeds, giving 24 seed–secret cases per evidence condition. A $d = 4$ study provides a smaller supporting instance.

Four support conditions separate information sufficiency from parameterized Recovery. Oracle support directly identifies the target. The non-oracle sufficient support is constructed without the direct target atom but has singleton version space under the known family. The two insufficient supports leave, respectively, exactly two secrets and a larger ambiguous set. For every support S, the radius is computed by exhaustive family enumeration,

$$r_S(k) = \max_{F_t \in \mathcal{V}(S)} d_{\text{long}}(F_s, F_t).$$

Because every distinct pair in this family has the same distance, singleton supports have $r_S(k) = 0$, while either insufficient support has $r_S(k) = D_{6,k}$. Thus the experiment can determine whether a failure remains possible from the supplied information before measuring neural execution.

k	$r_S = D_{6,k}$	wrong secret	two-secret Bayes	always zero	always one
1	0.031250	0.968750	0.984375	0.984375	0.015625
2	0.061035	0.938965	0.969482	0.968994	0.031006
4	0.116423	0.883577	0.941789	0.938950	0.061050
8	0.211853	0.788147	0.894073	0.881626	0.118374
16	0.351110	0.648890	0.824445	0.777265	0.222735
32	0.484172	0.515828	0.757914	0.604141	0.395859
44	0.505515	0.494485	0.747242	0.500110	0.499890
128	0.232064	0.767936	0.883968	0.133215	0.866785
160	0.148519	0.851481	0.925740	0.080481	0.919519

Table 2: Exact horizon-dependent references for $d = 6$. "Wrong secret" is $1 - D_{6,k}$. The two-secret Bayes column is $1 - D_{6,k}/2$ under equal posterior mass on the two remaining secrets. The constant baselines are exact under the theorem's uniform-block deployment distribution.

We report three distinct Recovery measurements. Support fit is accuracy on the presented episode. Secret Recovery requires the inferred secret to equal the held-out target. Local trigger-rule accuracy exhaustively evaluates all 2^6 one-block strings, rather than sampling deployment endpoints.

Support	Secret Recovery	Support fit	Local $k = 1$	Intermediate natural
Oracle	24/24	–	1.000000	1.0000
Sufficient, non-oracle	24/24	0.9948	1.000000	0.9883
Insufficient, $\lvert\mathcal{V}\rvert = 2$	0/24	–	0.984375	0.8075
More ambiguous	0/24	–	0.890625	0.6666

Table 3: Primary $d = 6$ Recovery and intermediate-window deployment results. Each condition contains 24 seed–secret cases. Natural accuracy is averaged over the four listed horizons under the theorem's deployment distribution.

The oracle arm has natural endpoint accuracy 1.0000 at every evaluated horizon. The sufficient non-oracle arm is approximately 1.0000 through $k = 16$, then attains $0.9750, 0.9788, 0.9938, 0.9701$ at

$k = 32, 44, 128, 160$, respectively. Its mean support fit is not exactly one: 23 of 24 cases fit every support example, while seed 2 with secret 42 fits 7 of 8 examples (0.875). Nevertheless, all 24 recover the secret and attain perfect exhaustive one-block recovery.

Under two-hypothesis support, neural behavior closely follows what remains identifiable. Across all nine evaluated horizons, its largest deviation from the two-secret Bayes reference in Table 2 is about 0.0099, at $k = 44$. Its mean natural accuracy in the sampled intermediate window is 0.8075, compared with the analytic reference 0.8059. At $k = 128$, $r_S = 0.2321$ and natural accuracy is 0.8783, while the always-one baseline is already 0.8668. At $k = 160$, these values are 0.1485, 0.9270, and 0.9195. Secret Recovery remains $0/24$: as the deployment criterion becomes less discriminating, endpoint accuracy rises without recovery of the hidden secret.

Natural accuracy is the deployment criterion $\Pr_{x\sim D_{\text{long}}}[\widehat{F}(x) = F_s(x)]$. We separately compute positive-class accuracy, negative-class accuracy, and their unweighted mean as a balanced diagnostic against class saturation. These diagnostics are not substitutes for the theorem's natural distribution. The analytic always-zero, always-one, and wrong-secret references are given in Table 2. At $k = 160$, the sufficient arm has positive-class accuracy approximately 0.9819 and negative-class accuracy approximately 0.8379, exposing a milder false-positive weakness that its natural accuracy partly masks.

One sufficient-support case shows a sharper execution limitation. For seed 0 and evaluation secret 1, natural accuracy is approximately 0.4023 at $k = 32$ and 0.4932 at $k = 44$. The same neural model and target achieve 1.0000 at every horizon under oracle support, while both support conditions recover the correct secret and exhaustive one-block rule. The failure is therefore not information insufficiency; support representation can still interact with parameterized Recovery and execution after $r_S = 0$.

The $d = 4$ experiment gives the same qualitative pattern without the localized sufficient-support failure: oracle and sufficient evidence recover the target, insufficient supports remain ambiguous, intermediate horizons separate the evidence conditions, and late-horizon insufficient-support accuracy approaches the always-one baseline while secret Recovery remains absent.

E.4 Endpoint-only active acquisition

The learner receives only endpoint answers $x \mapsto f^\star(x)$. MQ-only uses membership queries without target state labels. MQ+CE additionally receives an ideal counterexample and is reported only as a structural reference. Passive sampling draws endpoints without adaptation. Query cost is measured at the first budget whose mean long-horizon accuracy is at least 0.99.

Task	Acquisition process	Labels to criterion	Solved / censored
Random DFA	MQ-only	222.4 ± 86.7	5/5
Random DFA	MQ + ideal CE	130.2 ± 9.7	5/5
Random DFA	Passive	698.0 ± 382.7, solved only	4/5; $1 > 4096$
HMC	MQ-only	297.7 ± 69.1	3/3
HMC	MQ + ideal CE	175.0 ± 62.0	3/3
HMC	Passive	–	0/3; $3 > 4096$

Table 4: Endpoint-only acquisition. Costs are mean $\pm$ standard deviation among solved runs. The Random-DFA passive cost excludes one right-censored run and is not an unconditional expected cost. HMC passive has no solved-run cost; all three runs are right-censored above 4096 labels, where mean accuracy is about 0.452. MQ + ideal CE uses a stronger oracle and is not cost-equivalent to the endpoint-only arms.

The HMC active hypotheses recover complete behaviorally sufficient quotients and achieve 1.000 long-horizon accuracy. Under a guarded endpoint oracle, the MQ-only execution completes with zero latent-state accesses. Four deliberate latent-access probes fail, and static inspection finds zero latent-reaching lines on that execution path. The oracle-transition condition used elsewhere is separately labeled and is not part of MQ-only.

E.4.1 ACQUISITION SCALING

Random DFA nominal state counts are $N \in \{8, 16, 32, 64, 128, 256\}$. Machines are admitted only when the behavioral quotient obeys $|\bar{Q}| \geq 0.9N$. The resulting mean quotient sizes and success counts are:

Nominal N	8	16	32	64	128	256
Mean $\lvert\bar{Q}\rvert$	8.0	15.4	30.8	60.1	121.8	243.6
Passive RPNI	10/10	9/10	2/10	0/10	0/5	0/5
MQ-only	10/10	9/10	9/10	10/10	5/5	2/5
MQ + ideal CE	10/10	10/10	10/10	10/10	5/5	5/5

Table 5: Acquisition success as behavioral quotient size grows. A solve means reaching the declared 0.99 long-horizon criterion within budget.

At $N = 256$, the three MQ-only runs below the 0.99 solve threshold still attain 0.9602–0.9859 with complete transition functions. Descriptive log–log fits over the tested finite range give MQ-only query-cost exponents 1.3687 at machine level, with 95% confidence interval $[1.183, 1.555]$, and 1.2203 for scale means. Ideal-CE exponents are approximately 1.25–1.28. These fits are descriptive, not asymptotic laws, and the ideal-counterexample oracle is not cost-equivalent to endpoint-only acquisition.

E.5 ACQUISITION THRESHOLD SENSITIVITY AND ENDPOINT ACCOUNTING

Across solve thresholds $\tau \in \{0.95, 0.975, 0.99\}$, the qualitative active–passive ordering is stable, although exact solved fractions change at some scales. The verified counts at the scales used for the robustness comparison are:

Nominal N	Process	Seeds	$\tau = .95$	$\tau = .975$	$\tau = .99$
32	Passive	10	7/10	5/10	2/10
32	MQ-only	10	10/10	10/10	9/10
64	Passive	10	0/10	0/10	0/10
64	MQ-only	10	10/10	10/10	10/10
128	Passive	5	0/5	0/5	0/5
128	MQ-only	5	5/5	5/5	5/5
256	Passive	5	0/5	0/5	0/5
256	MQ-only	5	5/5	4/5	2/5
256	MQ + ideal CE	5	5/5	5/5	5/5

Table 6: Acquisition threshold sensitivity. Counts vary quantitatively with the solve criterion, but the large-scale active–passive ordering does not reverse.

MQ-only and passive sampling receive the same endpoint-label type $x \mapsto f^\star(x)$ and charge one cost unit per distinct labeled endpoint. Repeated active queries are cached, duplicate passive strings count once, and MQ-only receives no target-state metadata. MQ-only differs by selecting the next endpoint from previous answers. MQ + ideal CE additionally receives an ideal equivalence/counterexample oracle and is reported only as a stronger structural reference.

The primary passive RPNI implementation uses $\texttt{max_nodes} = \min(\texttt{budget}, 4000)$. An uncapped control uses budgets 4096 and 8192. At $N = 128$, its largest condition contains approximately 8184 distinct endpoint labels yet solves $0/3$ targets at the 0.99 criterion, with mean top-budget long-horizon accuracy approximately 0.7516. Thus the observed passive shortfall under this sampling process and learner is not explained solely by the node cap.

E.6 SAT DIAGNOSIS OF PASSIVE SAMPLES

The diagnostic asks whether a passive sample is itself ambiguous or whether the learner fails despite the absence of an alternative in the encoded class. For each cell, a SAT encoding searches

a fixed canonical reachable n-state Moore-machine class for a model that is consistent with every observed endpoint label and behaviorally inequivalent to the target. Transition, output, reachability, and sample-consistency variables are encoded exactly within this class. Canonical symmetry breaking selects one state naming; it does not establish equivalence to the unrestricted class of all machines with at most n states.

Diagnostic outcome	Cells
Learner fails; inequivalent consistent alternative found	47
Learner fails; no alternative found in encoded class	5
Learner succeeds; inequivalent consistent alternative found	2
Solver-undetermined at propagation cap	48

Table 7: Version-space diagnosis of the reported Random-DFA passive cells. A propagation-limit outcome is neither SAT nor UNSAT and is not included in a solved-cell proportion.

The first row diagnoses evidence ambiguity within the encoded class. The second diagnoses a learner/algorithm failure relative to that class. The third shows that ambiguity need not prevent a particular learner from selecting the target behavior. Behavioral inequivalence is the strongest terminology supported here: the long-horizon probe pool is not called an exact disagreement probability, so these results do not certify an exact ϵ-bad alternative. Exact product-automaton post-processing would be required for that stronger statement.

E.7 Identification enumerations

The additive and affine enumeration uses $p = 5$ and one symbol. All five additive rules and all 25 affine parameter pairs are enumerated. One additive witness leaves one rule. For affine updates, one witness or a repeated-source pair leaves five rules, while two distinct-source witnesses leave one.

The hidden-transducer enumeration fixes six states, two symbols, the initial state and endpoint readout, a finite family of local transition rules, and all endpoint observations through length four. Complete short endpoints leave worst-case long-horizon radius 0.8125; one third of relevant transition witnesses reduce it to 0.4375; the full relevant transition table reduces it to zero. These values are instance-specific finite calculations.

E.8 Hidden Modular Color

HMC has 16 latent states, four symbols, training length 16, and eight output classes. Endpoints reveal only the final class. Transition witnesses reveal selected triples $(q, a, \delta(q, a))$; local recovery evaluates the full target transition table.

Evidence	Achieved coverage	Local recovery	Seeds
Endpoint only	0.00	0.0469 ± 0.0128	3
Matched extra endpoints	0.00	0.0573 ± 0.0195	3
Transition witnesses	0.25	0.2813 ± 0.0221	3
Transition witnesses	0.50	0.516 ± 0.017	5
Transition witnesses	0.75	0.750 ± 0.000	5
Full transition	1.00	1.000 ± 0.000	3

Table 8: HMC Recovery by exposed transition coverage. Coverage reports the fraction of the declared transition table supplied during supervision and is distinct from LGID.

With full transition supervision and a supervised neural ϕ-head, all five seeds attain 1.000 transition recovery, per-step rollout state, final rollout state, neural readout, and end-to-end neural-readout rollout output at every normalized horizon $1, 2, 4, 8, 16$. At $16\times$, the direct endpoint output is about 0.117 versus chance 0.125.

L/L_{train}	Transition	Per-step state	Final state	Neural readout	Rollout output
1	1.000	1.000	1.000	1.000	1.000
2	1.000	1.000	1.000	1.000	1.000
4	1.000	1.000	1.000	1.000	1.000
8	1.000	1.000	1.000	1.000	1.000
16	1.000	1.000	1.000	1.000	1.000

Table 9: HMC full-transition neural-readout closure over five seeds.

E.9 General-K validation on learned HMC models

The validation population contains 560 partially recovered HMC checkpoints satisfying

$$0.2 < A_{\text{transition}} < 0.95.$$

No checkpoint is selected using its survival curve. The learned transition head predicts each (q, a) entry separately, so its symbolic transition table is reconstructed exactly. HMC evaluation draws each of the four symbols with $\nu(a \mid q) = 1/4$. For every checkpoint, define

$$K_{qq'} = \sum_a \frac{1}{4}\, \mathbf{1}\{\delta(q,a) = q',\ \widehat{\delta}(q,a) = q'\}.$$

The predicted quantity

$$S_L^K = e_{q_0}^\top K^L \mathbf{1}$$

is compared with empirical exact no-error trajectory survival at every integer horizon $L = 1, \ldots, 256$. The comparison counts a trajectory as failed after its first state divergence.

Survival predictor	Short-horizon MAE
State-dependent operator $e_{q_0}^\top K^L \mathbf{1}$	6.25×10^{-4}
Target-visitation scalar $\prod_t p_t$	9.98×10^{-3}
Unweighted scalar $A_{\text{transition}}^L$	1.24×10^{-1}

Table 10: Prediction error across 560 partially recovered HMC checkpoints over the reported short-horizon comparison band. The operator is about $16\times$ more accurate than the target-visitation scalar and about $200\times$ more accurate than the table-mean scalar.

The unweighted scalar loses both state visitation and error location. The visitation-aware scalar uses the correct target occupancy π_t through

$$p_t = \sum_q \pi_t(q) c(q), \qquad \widehat{S}_L^{\text{visit}} = \prod_{t=0}^{L-1} p_t,$$

but it still does not condition occupancy on survival. The exact operator propagates the surviving mass and therefore uses $\widetilde{\pi}_t$ implicitly. Under uniform symbol input, the target HMC transition chain is doubly stochastic, so its stationary state distribution is uniform. The stationary visitation-weighted scalar consequently equals the unweighted transition-table scalar and is not a distinct comparator for this task. Proposition C.3 shows that the remaining factor gap is exactly the covariance between historical survival propensity and next-step survival; Corollary C.4 bounds it by survival heterogeneity times the induced occupancy shift.

There are also 299 compatible full-Recovery checkpoints with $A_{\text{transition}} = 1$. For these checkpoints, $S_L^K = 1$ exactly at every evaluated horizon.

Pair-state resynchronization. At $L = 256$, the partially recovered population has mean exact no-error survival

$$S_{256} \approx 1.24 \times 10^{-10},$$

whereas current synchronization is

$$C_{256} \approx 0.0625 = 1/16.$$

Thus almost no trajectory remains error-free, while the pair process approaches the uniform 16-state synchronization level through repeated divergence and resynchronization. No-error survival and current synchronization are qualitatively different long-horizon quantities.

Time-varying stress tests. For nonstationary inputs, the exact object is the ordered product $K_0K_1\cdots K_{L-1}$. Two-phase and gradual schedules verify this product directly. In the tested schedules, the target-visitation scalar differs by roughly 1%–8%, so nonstationarity changes the exact object without implying a universal large scalar gap; the approximation quality depends on the schedule.

Finite-horizon and spectral summaries. The exact finite-horizon identity applies to every checkpoint. Of the 560 partial-model kernels, 95 (approximately 17%) are primitive and additionally admit the single-radius Perron–Frobenius asymptotic in Theorem 3.7. The empirical comparison uses the finite-horizon operator for the full population.

E.10 Controlled one-error HMC tables

The HMC transition table contains $16 \times 4 = 64$ state–symbol entries. For each entry, replace its target next state by each of the other 15 states while leaving the remaining 63 transitions correct. This enumerates $64 \times 15 = 960$ controlled tables, each with exactly $A_{\text{transition}} = 63/64$. Exact operator evaluation gives

$$S_{16} \in [0.5933, 0.8743], \qquad S_{256} \in [0.0110, 0.0706].$$

The local error count is identical throughout the enumeration. The survival range is therefore caused by which transition is wrong and how often surviving deployment paths encounter it.

E.11 Matched Local Recovery, Different Finite-Horizon Survival

To separate error placement from the amount of average local Recovery, we bucket the same learned checkpoints by their exact transition-table accuracy. For a reconstructed kernel K, we report both the original- q_0-anchored survival

$$S_L^{q_0} = e_{q_0}^\top K^L \mathbf{1}$$

and the start-state-averaged robustness quantity

$$\overline{S}_L = |Q|^{-1}\mathbf{1}^\top K^L \mathbf{1}.$$

The latter prevents a single designated start state from obscuring differences elsewhere in the learned table; neither quantity treats checkpoints from one training trajectory as independent replicates.

Exact-accuracy bucket	$A_{\text{transition}} = 0.296875$	$A_{\text{transition}} = 0.265625$
Checkpoints	53	29
Distinct learned tables	20	15
Experiment families	8	7
Methods	23	24
Learner seeds	5	5

Table 11: Support of two exact local-Recovery buckets. Checkpoints may be correlated within training runs; the counts describe the learned population and its structural diversity, not independent replication.

In the $A_{\text{transition}} = 0.296875$ bucket, two of the 53 checkpoints have exactly zero start-state-averaged survival. Among the nonzero checkpoints, $\overline{S}_{16}$ ranges from 3.783×10^{-10} to 7.026×10^{-5}, with median 5.767×10^{-5}, while $\overline{S}_{32}$ ranges from 8.809×10^{-20} to 1.787×10^{-8}, with median 1.115×10^{-8}. The range and median explicitly exclude exact zeros. Under the original q_0-anchored criterion, 37 of 53 checkpoints have exactly zero $S_{16}^{q_0}$, and the same 37 have zero $S_{32}^{q_0}$. Thus the zero/nonzero distinction is not an artifact of reporting a large ratio between tiny values.

A matched learned pair in this bucket makes the contrast concrete. Both have exactly $A_{\text{transition}} = 0.296875$, but their start-state-averaged values are

$$(\overline{S}_{16}, \overline{S}_{32}) = (0, 0) \quad \text{and} \quad (5.767 \times 10^{-5}, 1.787 \times 10^{-8}).$$

The nonzero $\overline{S}_{16}$ is approximately the bucket median, whereas its $\overline{S}_{32}$ is the bucket maximum. The second exact-accuracy bucket, $A_{\text{transition}} = 0.265625$, contains 29 checkpoints and 15 distinct

learned tables and again spans many orders of magnitude in finite- horizon survival. The conclusion is structural: identical average local Recovery can place errors on qualitatively different surviving trajectories, so it does not determine finite-horizon Composition reliability.

E.12 FOUR-STAGE HMC FINGERPRINT

All five conditions use the same HMC task, Transformer architecture, optimizer, transition/readout rollout pathway, and horizon-256 endpoint metric. The controlled perturbation targets one diagnostic knob: the admissible channel, acquired subset, optimizer exposure, or location of residual local errors. The neural parameterization remains shared, so a stage-matched intervention can alter other Recovery metrics. Fifteen seeds are run per condition. For declared thresholds $(\epsilon, \tau_{\mathrm{local}}, \tau_{\mathrm{comp}})$, the Composition label means that the channel and sample are sufficient and local fit exceeds τ_{local}, but horizon-256 survival is below τ_{comp}; it does not assert exact local Recovery.

For a channel O, let r_O be its worst-case long-horizon version-space radius. Let r_S be the radius after the evidence actually acquired, A_{seen} fit on presented constraints, $A_{\mathrm{transition}}$ accuracy on the full transition table, and S^K_{256} exact operator survival for the reconstructed table.

Condition	r_O	r_S	A_{seen}	$A_{\mathrm{transition}}$	S^K_{256}
Healthy	0	0	≈ 1	≈ 1	≈ 1
Identification-limited	≈ 0.868	–	1.000	–	–
Acquisition-limited	0	≈ 0.875	1.000	–	–
Recovery-limited	0	0	–	≈ 0.848	–
Composition-limited	0	0	–	0.984375	≈ 0.015

Table 12: Stage metrics for the controlled HMC fingerprint. Dashes denote metrics not needed to locate the first criterion-relative bottleneck, rather than imputed values. The signature is stable over 15 seeds per condition.

The four failure conditions have horizon-256 learned-rollout endpoint accuracy between 0.1235 and 0.1364, for a total span of 0.0128. HMC's eight-class learned-rollout score is effectively bimodal in this sweep: incorrect tables concentrate near the chance floor $1/8$, while exact tables score 1.0. The comparison therefore matches the four failures near the floor.

Diagnosed stage	Intervention	Outcome
Identification	Restore decisive witness	Solved
Identification	Same channel, 2400 more updates	$0/9$ repaired
Acquisition	Acquire missing evidence	$9/9$ solved
Acquisition	Same acquired set, 2400 more updates	$0/9$ repaired
Recovery	Complete evidence, increase exposure	$9/9$ solved
Composition	Ordinary full-objective training	Repaired

Table 13: Stage-matched fingerprint interventions. Information-limited conditions do not respond to more optimization on unchanged evidence; the fitting and high-impact-residual conditions do.

A targeted update on one Composition pair is not an independent table edit: because state and symbol embeddings are shared, the update changes other rows and reduces $A_{\mathrm{transition}}$ from approximately 0.984 to 0.810. This negative arm connects the fingerprint to both the shared-parameter Recovery model and the state-dependent Composition analysis.

E.13 ENDPOINT-ONLY ARCHITECTURE COMPARISON

The Looped HMC run does not fit training: accuracy is about 0.186 at 600 updates and 0.162 at 2400. Its long-horizon result is therefore an optimization failure and does not isolate whether looping helps or hurts length generalization.

Architecture	HMC 16×	Random DFA 16×
Standard Transformer	0.099	0.491
Looped Transformer	0.113	0.444
GRU	0.120	0.605
Chance	0.125	0.500

Table 14: Endpoint-only long-horizon accuracy for the tested configurations. The task family and horizons match the acquisition study, while architecture runs fix task seed 0 and vary training randomness.

E.14 Exact addition

Addition uses base-10 digits, two carry states, training length 8, five seeds, and evaluated lengths $8, 16, 32, 64, 128$. Full transition supervision gives exact transition and rollout-state recovery. The true task emitter gives exact sums at every tested horizon. The learned emitter has per-digit accuracy approximately 0.993, while exact-sequence accuracy is:

L/L_{train}	1	2	4	8	16
Learned-emitter exact sequence	0.936	0.904	0.807	0.653	0.418
0.993^L reference	0.945	0.894	0.799	0.638	0.407
Task-emitter exact sequence	1.000	1.000	1.000	1.000	1.000

Table 15: Addition Composition with exact state rollout. The independent reference is within 0.021 of the observed learned-emitter sequence accuracy.

The partial-carry condition is secondary because it mixes Recovery and Composition. At half coverage, carry recovery is 0.877 ± 0.022, and exact sums fall

$$0.358 \to 0.127 \to 0.019 \to 0.0016 \to 0$$

over the same normalized horizons.

Across carry states, learned-emitter accuracy differs by only about 0.0139 on average. A state-dependent emission kernel therefore improves prediction only marginally relative to the homogeneous p^L description, so Addition serves as the scalar special case in the main text.

E.15 Bounded-Dyck Recovery under normalized exposure

The scaling study uses

$$N \in \{32, 64, 128, 256, 512, 1024, 2048\}$$

bounded-stack states. The learner is fixed at $d_{\text{model}} = 96$, two layers, four heads, and $d_{\text{ff}} = 256$. Full transition supervision contains $M_N = |Q||\Sigma|$ state–symbol pairs, and evaluation uses this same declared table. The reported metric is therefore full-table fit, not unseen-transition generalization. All scaling runs execute CPU-only.

Define transition-table exposure by

$$E = \frac{\text{updates} \times \text{batch size}}{M_N}.$$

Comparing fixed E holds constant the number of optimization examples per supervised transition.

Systems from 32 through 2048 states all remain near 0.51 full-table fit at $E = 4$, so low exposure can produce severe underfitting even in the smallest tested system. At $E = 16$, supervised full-table Recovery is exact through 512 states and nearly exact at 1024. The 2048-state condition reaches 1.000 on 3/3 seeds when exposure is increased from $E = 16$ to $E = 32$. Thus raw update comparisons conflate table growth with declining visits; normalized exposure isolates this axis while leaving architecture-dependent rank and conditioning as separate factors.

N	$E = 4$	$E = 16$	$E = 32$
32	≈ 0.51	1.000	–
64	≈ 0.51	1.000	–
128	≈ 0.51	1.000	–
256	≈ 0.51	1.000	–
512	≈ 0.51	1.000	–
1024	≈ 0.51	0.99992	–
2048	≈ 0.51	0.97636 ± 0.00924	1.000

Table 16: Bounded-Dyck full-transition-table fit under normalized exposure. At $E = 16$, all three 1024-state seeds exceed 0.99, while no 2048-state seed does; increasing only the 2048-state condition to $E = 32$ restores exact fit on 3/3 seeds.

Sink-collapse mechanism. Across the 36 low-exposure checkpoints at $E = 4$, the aggregate full-table score hides a strongly asymmetric predictor:

Metric	Mean
A_{all}	0.528
$R_{\text{pred sink}}$	0.974
A_{sink}	0.9995
A_{nonsink}	0.0359
$A_{\text{balanced}} = \frac{1}{2}(A_{\text{sink}} + A_{\text{nonsink}})$	0.518

The most extreme condition is $N = 32, E = 4$: every table entry is predicted as the sink, non-sink accuracy is zero, and overall accuracy exactly equals the always-sink baseline. The degradation is therefore a shared majority shortcut rather than uniform corruption. Increasing exposure restores the non-sink corrections, matching the qualitative timescale separation of Proposition 3.6.

Fixed-update comparison. For the 512-state system, full supervision provides $512 \times 4 = 2048$ transition pairs. At width 96, fit rises from 0.5535 ± 0.0186 after 500 updates to 0.99984 ± 0.00023 after 2000; width 192 rises from 0.5885 ± 0.0299 to 1.000. At 500 updates, the sequence model predicts the sink state for roughly 91–97% of pairs; always predicting the sink gives about 0.502. Exact lookup attains 1.000, and a standalone MLP attains 0.9995. This fixed-update comparison illustrates why raw update count confounds state-space size with exposure to the growing table.

E.16 Secondary Recovery diagnostics

Task or condition	Evidence	Local recovery	Seeds
Random DFA	Endpoint only	0.133 ± 0.041	5
Random DFA	Full transition	1.000 ± 0.000	5
Bounded Dyck, 32 states	Endpoint only	0.019 ± 0.014	5
Bounded Dyck, 32 states	Full transition	1.000 ± 0.000	5
Renamed DFA, state names	Endpoint only / full	$0.047/0.063$	5
Renamed DFA, symbol names	Endpoint only / full	$0.078/0.313$	5

Table 17: Secondary Recovery diagnostics. Renaming measures transfer across representations, not recovery within the observed coordinate system.

Nominal coverage	Witness noise	Clean coverage	Local recovery
0.50	0.00	0.500	0.516
0.50	0.20	0.432	0.448
1.00	0.00	1.000	1.000
1.00	0.20	0.797	0.797

Table 18: HMC recovery with corrupted transition witnesses (three seeds). Effective clean coverage tracks local recovery in this diagnostic.

E.17 A SHORTCUT OUTSIDE THE LOCAL-TABLE CLASS

The fixed-table upper bound in Theorem D.4 depends on the reference class. Let $f(x_1, x_2) = x_1 \vee x_2$. Observing $(0,0) \mapsto 0$, $(1,0) \mapsto 1$, and $(0,1) \mapsto 1$ removes alternatives that differ on one single-variable change, but XOR remains consistent and fails at $(1,1)$. Local coverage cannot eliminate a compositional shortcut that the class allows only at a joint input.

F VERIFIED EXPERIMENTAL CONFIGURATION

The experimental audit records 136 hyperparameter values across six experiment families, with no unavailable fields. Each value is traced to an explicit configuration, a proven code default, or saved checkpoint metadata. The tables below report the verified values supplied for the main HMC and bounded-Dyck neural studies; no unspecified value is inferred.

HMC Transformer field	Verified value
d_{model}	128
Transformer layers	2
Attention heads	4
d_{ff}	256
Positional encoding	Learned
Dropout	0.0
Optimizer	AdamW
Learning rate	10^{-3}
Adam betas	$(0.9, 0.999)$
Weight decay	0
Maximum modeled length	256

Table 19: Verified HMC Transformer configuration.

HMC architecture	Parameter count
Standard Transformer	376,880
Looped Transformer	376,880
GRU	199,688

Table 20: Verified HMC architecture parameter counts.

Bounded-Dyck scaling field	Verified value
d_{model}	96
Transformer layers	2
Attention heads	4
d_{ff}	256
State counts	$32, 64, 128, 256, 512, 1024, 2048$
Exposure	$E = UB/M$
Execution hardware	CPU only

Table 21: Verified bounded-Dyck scaling configuration.

Acquisition aggregates vary task instances and learner randomness; the architecture comparison fixes task seed 0 and varies training randomness. Censored passive runs, distinct-query accounting, threshold definitions, and the stronger ideal-counterexample oracle are reported in Appendix E.5.